Dependencies in evidential reports: The case for informational advantages.

Toby D. Pilditch[1*] (t.pilditch@ucl.ac.uk), Ulrike Hahn[2] (u.hahn@bbk.ac.uk), Norman Fenton[3] (n.fenton@qmul.ac.uk), and David Lagnado[1] (d.lagnado@ucl.ac.uk)

[1]University College London, Gower Street, London WC1E 6BT, UK;

[2]Birkbeck University of London, Malet Street, Bloomsbury, London, WC1E 7HX, UK;

& [3]Queen Mary University of London, Mile End Road, London, E1 4NS, UK

*Corresponding Author: Correspondence should be addressed to Toby D. Pilditch, 26 Bedford Way, London, WC1H 0AP, UK. Electronic mail can be sent to t.pilditch@ucl.ac.uk.

Declarations of interest: None.

Abstract

Whether assessing the accuracy of expert forecasting, the pros and cons of group communication, or the value of evidence in diagnostic or predictive reasoning, dependencies between experts, group members, or evidence have traditionally been seen as a form of *redundancy*. We demonstrate that this conception of dependence conflates the *structure* of a dependency network, and the *observations* across this network. By disentangling these two elements we show, via mathematical proof and specific examples, that there are cases where dependencies yield an informational *advantage* over independence. More precisely, when a structural dependency exists, but observations are either partial or contradicting, these observations provide more support to a hypothesis than when this structural dependency does not exist, *ceterus paribus*. Furthermore, we show that lay reasoners endorse sufficient assumptions underpinning these advantageous structures yet fail to appreciate their implications for probability judgements and belief revision.

Keywords: evidential reasoning; probabilistic reasoning; dependence; reliability; belief updating

1. Introduction

"Surely, Mr. Lincoln," said I, "that is a strong corroboration of the news I bring you." He smiled and shook his head. "That is exactly why I was about you about names. If different persons, not knowing of each other's work, have been pursuing separate clews that led to the same result, why then it shows there may be something in it. But if this is only the same story, filtered through two channels, and reaching me in two ways, then that don't make it any stronger. Don't you see?"

William Seward (1877)

Consider the following scenario: a plane has crashed, and you must determine whether it was sabotage. You await the crash site reports from two investigators, Bailey and Campbell. They have separately assessed the various pieces of wreckage before leaving to write up their conclusions. Both investigators are equally accurate in their conclusions, seldom making mistakes. Now consider two alternative cases:

i. Bailey provides a report in which she concludes the plane was sabotaged, but she has also seen Campbell's report, in which Campbell likewise concluded that the plane was sabotaged.
ii. Bailey provides a report in which she concludes the plane was sabotaged, based on her assessment alone. Campbell then separately provides a report (based on his assessment alone), likewise concluding that the plane was sabotaged.

Here, i) is a case of corroborating reports with a directional dependence from Campbell to Bailey (i.e., Bailey has seen Campbell's report, thus Bailey's report may *depend* upon Campbell's, but not vice-versa), and ii) is a case of corroborating reports coming from independent sources. Given the two reports in each case (*ceterus paribus*), it would be right to conclude that more support for the conclusion that the plane was sabotaged is provided in the independent case. Indeed this is how dependencies have traditionally been viewed (see e.g., Nisbett & Ross, 1980): namely, as a

compromising influence that makes additional evidence redundant. However, now consider the same scenario, with two slight alterations:

1. Bailey reports to you the plane was sabotaged, having seen Campbell's report, but *you do not know what Campbell concluded* (case i), versus you only know Bailey's independent conclusion of sabotage (case ii).
2. Bailey reports to you the plane was sabotaged, having seen Campbell's report, but *you know that Campbell concluded the opposite* (case i), versus you only know that Bailey and Campbell have independently provided contradictory conclusions (case ii).

1) is an instance of *partial* information, and 2) an instance of *contradicting* information. In both these instances, it is less clear whether case i) or ii) provides more support for the sabotage hypothesis. In the present paper, we demonstrate that for partial or contradicting information, the dependent case (i) is, in fact, superior (i.e., there is a dependency advantage, in that more evidential support is provided to the hypothesis when a report is the result of a structural dependency (i) than when independent (ii)) – at least given reasonable assumptions.

1.1. Dependence in Evidential Reasoning

Dependence has been considered in many forms. For groups, it has often been conceptualised as the degree of correlation between group member judgments, including juries (Berg, 1993; 1994; Ladha, 1995) and voting populations (Einhorn, Hogarth, & Klempner, 1977; Hogarth, 1978; Jönsson, Hahn, & Olsson, 2015; Hahn & Hornikx, 2016; Hahn, von Sydow, & Merdes, 2019). In the case of evidential reasoning, there are three primary forms of dependence: dependence as a shared background or reliability (e.g., economists found to have been educated in the same (weak) school; see Madsen, Hahn, & Pilditch, 2018; but also Bovens & Hartmann, 2003), dependence as direct information flow between sources (the focus of the present paper, but see

also instances of conferring between witnesses or suspects; Wagenaar, Van Koppen, & Crombag, 1993), and dependence as shared evidence (Schum, 1994) – as outlined by Lincoln above[1].

Lincoln's supposition on the nature of dependence fits with the general consensus across multiple literatures: the evidential and causal reasoning literature (Pearl, 1988; 2009; Schum, 1994; Rehder, 2014), expert forecasting (Hogarth, 1989; Soll, 1999) and the literatures on applied domains of reasoning such as medicine (Kononenko, 1993), law and forensics (Wagenaar, et al., 1993; Bex & Prakken, 2004; Dawid & Evett, 1997; Schum, 1994), risk analysis (Smith, Ryan, & Evans, 1992; Clemen & Winkler, 1999), and intelligence analysis (Spellman, 2011). This general consensus – that dependence is inferior to independence – has been based on instances of *corroboration*. Put another way, the issue of whether multiple items of evidence, reports, votes, or judgments are to some degree dependent is only raised *given they have been observed saying the same thing*. This is not surprising: contradictory reports might themselves be seen as evidence *against* dependence; and partial information (only observing a subset of the reports) may often leave the existence of a dependency unclear. From that perspective, it makes sense to formally treat dependence solely as a matter of correlations between outcomes, such as the content of reports, without consideration of how those outcomes come about; and this has, indeed, been the common strategy in the literature on collective judgment (see e.g., Berg, 1993; Ladha, 1995; Soll, 1999; Hahn & Hornikx, 2016; Hahn, et al., 2018; but for a discussion of the general difficulties of assessing dependencies, see Clemen, Fischer, & Winkler, 2000). However, as we show in this paper, this is not enough.

A fully adequate treatment of the normative implications of dependence requires consideration of *structure*; that is, explicit modelling of the connections between pieces of evidence that give rise to their specific content. To illustrate the varying impact of dependence,

[1] In fact, the example of Bailey and Campbell can be considered an instance where there is *shared evidence* across cases. However, we highlight that the key comparison remains the presence or absence of a *direct* dependence between the investigators.

and more generally model the integration of evidence, we turn to Bayesian Networks (BNs). These provide graphical representations of evidence-hypothesis structure, based on probabilistic dependency relations between variables (Pearl, 1988). In general, the Bayesian framework provides a characterization of optimal inference in the sense that Bayesian inference minimises the inaccuracies of one's beliefs (Pettigrew, 2016). In that sense, it provides a normative characterisation of what reasoners *should* ideally do. BNs in turn are a tool for simplifying the required Bayesian computations by exploiting dependence relations between variables. Hence BNs not only support optimal inferences when reasoning under uncertainty (Pearl, 2009), but also allow one to capture, by design, the impact of (at least certain types of) dependencies on complex inferences (see Schum, 1994).

The generic case of a direct dependency is illustrated in Fig. 1b, where a hypothesis H (e.g., 'sabotage') is informed by two sources, S_B (Bailey) and S_C (Campbell), who provide a report to the effect that the claim at issue (sabotage) is either true or false. The difference between case a) and b) is shown by the dashed arrow from S_C to S_B. This represents the presence of the directional dependency (Bailey receiving information from Campbell).

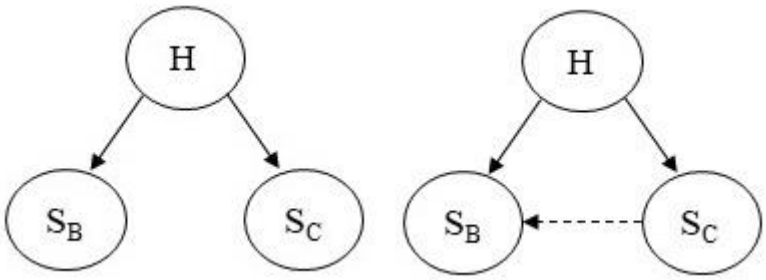

Figure 1a and 1b. Graphical representation of a hypothesis (H) with two sources of evidence (S_B, S_C) informing upon it. Fig. 1a (left) is the independent case, with the two sources reporting entirely separately. Fig. 1b illustrates the possible dependency from S_C to S_B (dashed line). The presence of a directed arrow indicates that the receiving node (S_B) is (directly) probabilistically dependent on the node at the arrow's origin (S_C).

Underpinning this graphical representation are a set of conditional probability tables (CPTs) associated with each node. These probabilities characterize precisely the variables and relations

between them: in the case of Fig. 1, the CPT associated with H will contain the prior probability of H (conditioned on any background information); the CPT associated with S_C contains the conditional probabilities of S_C given both H and not H; and the CPT associated with S_B is either identical to S_C (Fig. 1a, independent case) or contains the conditional probabilities of observing the different possible states of S_B as a function of the states of both S_C and H (Fig. 1b, dependent case). The graph and CPTs (which collectively define the BN) can then be used to calculate, via Bayes' rule, the influence of sources on the likelihood of the hypothesis being true.

Consequently, this formal framework is ideal for representing both general constraints on sources (such that more reliable[2] sources exert greater influence on the hypothesis) and for representing dependencies between sources. This then allows us to determine how rational agents *should* respond to the different cases outlined above.

1.2. Advantageous Dependence

There are many possible considerations when determining the nature of a dependency. For instance, to what degree does background information play a role in message passing between sources? Or in the accurate detection of the hypothesis in the first place? Are sources aware of their own or their fellow source's reliability? Such considerations are highly context-dependent, and speak to the difficulty of extricating a general rule of dependencies from the background information that surrounds it (Schum, 1994). However, here we seek to demonstrate that:

1. There exist cases that lead to an advantageous effect of a dependency (relative to an independent equivalent, *ceteris paribus*), resting on reasonable assumptions.

[2] In the present paper, reliability is taken to be synonymous with *accuracy*, i.e., a higher reliability source is less error-prone (for other Bayesian network representations of reliability see e.g., Fenton, Neil, & Lagnado, 2013; Hahn, Harris, & Corner, 2016).

2. When faced with such cases, participants endorse these assumptions, but do not infer the consequence (that the dependent structure is more advantageous).

1.2.1. Where is Dependence Advantageous?

Crucial to understanding where and when dependency advantages exist in cases such as the examples above, is appreciation of the impact of structure on them. What does this mean? It means appreciating that while there is a *directional dependency* such that Bailey's report is influenced by Campbell's, Campbell is *not* influenced by Bailey (see Fig. 1a vs 1b). The evidential value of Campbell's report is the same across the dependent and independent case. Whether or not there is a dependency advantage or disadvantage thus rests on the nature of Campbell's impact on Bailey. If seeing Campbell's report makes Bailey's report "better", then the dependency will be *advantageous*.

Crucially, the normative Bayesian framework makes clear that the evidential value (or 'diagnosticity') of a piece of evidence depends on the relationship between two cases: how likely that evidence would be if the hypothesis were true *and* how likely it would be if the hypothesis were false. Likewise, whether Bailey's report has greater evidential value when it depends (in part) on Campbell's depends both on the increase in accuracy when Campbell's report is true *and* the decrease in accuracy when Campbell misreports.

In the case of corroboration, this boils down simply to whether the potential benefits outweigh the costs to Bailey's accuracy: if Campbell's report increases accuracy more (when true) than it decreases accuracy (when false), then the two corroborating reports will provide stronger evidence in the dependent than the independent case. In other words, there will be a dependency advantage (see Appendix A and C for formal details, and Appendix B for visualisation).

In the case of contradictory reports, the evidence points in opposite directions. Where Bailey and Campbell are otherwise equally competent, their reports will simply cancel out in the

independent case. If the dependency weakens the evidential value of Bailey's testimony, however, then Bailey's conflicting report will counteract Campbell's to a lesser degree. So a dependency *disadvantage* for *corroboration* can become a *contradiction advantage*: the conflicting reports provide stronger evidence for the hypothesis in the dependent than independent case (see Appendix A and C for formal details).

In the partial information case, finally, we do not know, as it were, *which world we are in*: the one where Campbell's report corroborates Bailey's or the one where Campbell contradicts her. As a result, the benefits (or costs) of the dependency are based on the expectation over these two cases (see Appendix C for formal details).

These basic considerations make clear not only that dependency may, in the right circumstances, be beneficial, but also that the same circumstances can give rise to complex patterns of advantage/disadvantage across the corroboration, contradiction, and partial information cases.

To illustrate this further, Fig. 2 takes the two cases of the sabotage scenario, represents them in a BN fitted with parameters in line with two assumptions: sources are generally reliable (assumption 1; in this case error rates of independent sources are 20%), and dependencies reduce recipient error rates when correct information is provided by the sender (assumption 2; 20% error probability decreases to 10%). The left versus right hand columns represent dependence versus independence, and different rows represent different information states from initial state (top row) through the partial information case (middle row t_1) through to the contradicting information case (row t_2) stages.

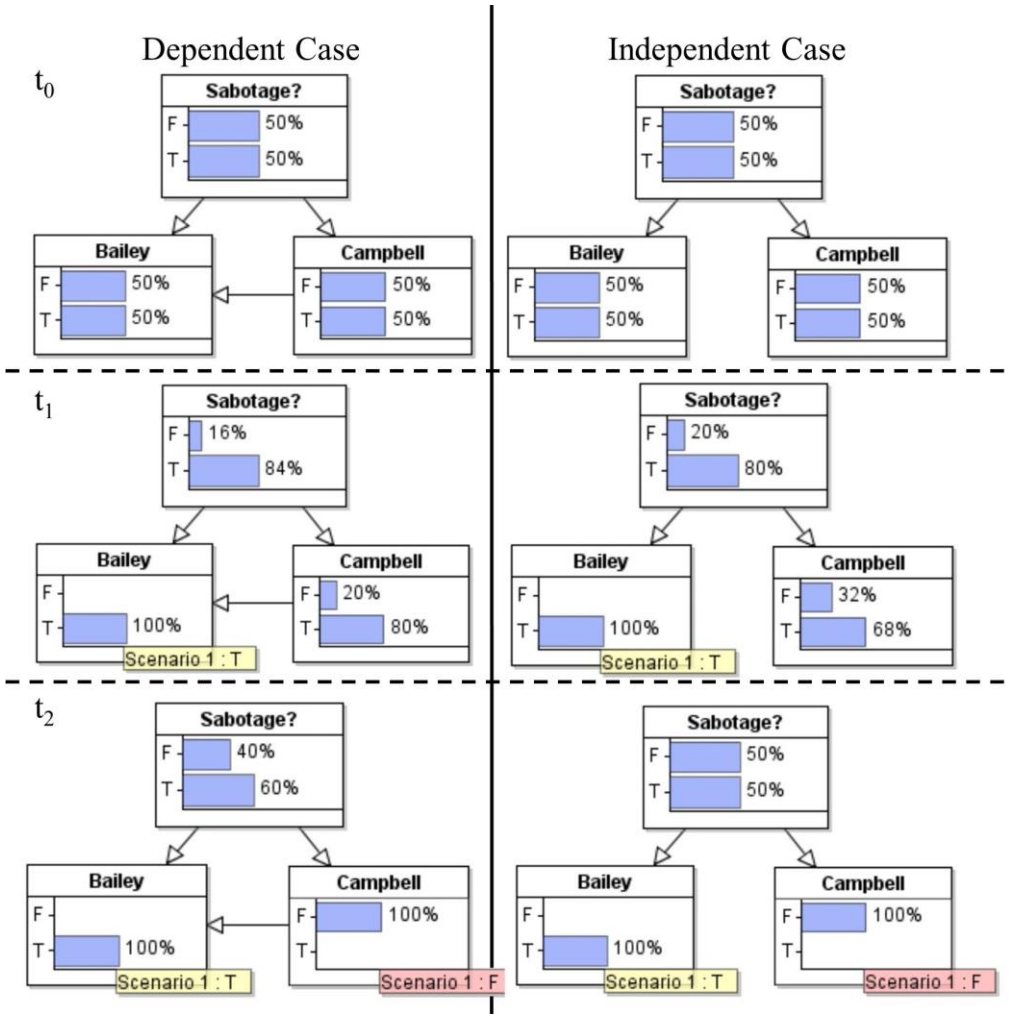

Figure 2. Bayesian Network illustrations of sabotage scenario, with independent (right-hand column) and dependent (left-hand column) cases. The top row (t_0) represents the sabotage scenario before any reports are observed. "Sabotage" is the underlying hypothesis at issue, and the other two variables are Bailey's and Campbell's reports. Blue bars and percentages indicate the respective probabilities for the possible states of those variables: e.g., the prior probability of sabotage (sabotage = T) is 50%; likewise, there is maximal uncertainty on whether Bailey and Campbell will report that there was sabotage (T) or that there was not (F). The middle row (t_1) represents a partial information state: we have observed Bailey's report (T) indicating sabotage, but have not yet observed Campbell's report. The new percentages indicate the updated probabilities in light of Bailey's report. Finally, in the bottom-row (t_2) we have now also observed Campbell's contradicting report (F), indicating it was not sabotage. Assumptions: P(Sabotage) = .5; Independent source error rates (P(T|¬Sabotage), P(F|Sabotage)) are all equal to .2; a dependency (Campbell to Bailey, left-hand column) entails a conditional halving (when Campbell is correct) and doubling (when Campbell is incorrect) of the (standard independent) .2 error rates of Bailey.[3]

[3] Figure created using the AgenaRisk Bayesian Network software (AgenaRisk, 2018).

What types of advantage/disadvantage exist in each case (corroboration, contradiction, partial information) will depend on the specific probabilities involved. However, the two conditions just outlined point to a particularly interesting region of the probability space. Where sources are more accurate than not, and where the additional information provided by Campbell helps, there will be broad regions of the parameter space where there are dependency advantages in the partial information case and in the contradictory information case.

We present a formal proof to illustrate dependency advantages for these two cases in the appendix. This proof demonstrates a dependency advantage *for the partial case*[4] in the sense that:

$$P(H|S_{B\text{-}Dep}) > P(H|S_{B\text{-}Ind}) \quad \text{(P1)}$$

In words, evidence provided by S_B will lead to a greater degree of belief in H (starting from the same prior, P(H)) in the dependent model.

It is necessary to highlight the assumptions that underpin these advantages, both to make the theoretical point that such advantages can *reasonably* occur, but further to identify the components to be endorsed (or not) by lay reasoners. We additionally note at this point that it remains possible that other assumptions (not detailed here) may also exist that can produce dependency advantages, but for the present case we use the worked through assumptions and proof to show this effect in principle. These two assumptions are:

1. "Sending" sources are generally accurate in the minimal sense that they more often provide correct information than not. More formally:

$$P(S_C|H) > 0.5 \quad \text{(A1)}$$

2. "Recipient" sources are assisted by the provision of correct information from the "sender":

[4] We note that in all the following probability formula, for a variable X (e.g., H, S_C, S_B, etc.), we take X to mean "X is true", and ¬X to mean "X is false".

$$P(S_{B\text{-}Dep}|H, S_C) > P(S_{B\text{-}Ind}|H) \quad (A2)$$

In words, these assumptions mean that sending sources are more likely to be accurate than inaccurate (i.e., be more likely to say "H is True" than "H is False" when H is in fact true, and more likely to say "H is False" than "H is True" when H is in fact false; assumption 1), and that a source evaluating first-hand evidence is less likely to make a mistake when also provided with access to *correct information* from another source (assumption 2). Arguably these assumptions are met *the majority of the time* in human communication: were they not, communication would not be worthwhile and it would be difficult to see why human beings evolved and continue to sustain such elaborate (and costly) communication. Hence, we argue these assumptions are reasonable approximations of dependency scenarios in both lay and professional contexts.

Finally, as the starting accuracy of sources drops close to the 0.5 threshold, a third assumption is needed to guarantee the dependency advantage, namely that:

3. "Recipient" sources should not be misled to a greater degree by incorrect information from a sender than they are assisted by correct information from a sender:

$$(P(S_{B\text{-}Dep}|H, S_C) - P(S_{B\text{-}Ind}|H)) > (P(S_{B\text{-}Ind}|H) - P(S_{B\text{-}Dep}|H, \neg S_C)). \quad (A3)$$

For the interested reader, the full functional form of these assumptions is detailed in Appendix A, and the mathematical proof demonstrating the dependency advantages that follow from them are contained in Appendix C, along with visualisations in B. To further illustrate these assumptions, consider again the case of Investigators Bailey, Campbell, the crashed plane, and the numerical example of Fig. 2:

Each investigator looks through the wreckage, noting details and features (e.g., state of control surfaces, fuel lines, engine, etc.), collating these to form their respective individual assessment of sabotage. In this process, there will be differences between Bailey and Campbell

in terms of a) the clues perceived, and b) the manner in which these clues are integrated to form a judgment, for example, Bailey and Campbell may have complementary expertise.

Now, let us first consider Bailey in the independent versus dependent case. In the former, she only has access to the clues she has gathered, and the way she has marshalled those clues to form her own judgment. In the latter, she not only has access to the above, but also has access to Campbell's judgment. Next, we consider the two critical assumptions. We know that Campbell generally provides accurate information (assumption 1). We also know that given accurate information, Bailey is less likely to make a mistake (and thus provide more support to the hypothesis) than if she did not have such information (assumption 2). One might imagine, for example, that Campbell accurately observes clues (or provides sound arguments) that point Bailey in the correct direction (e.g., Campbell catches a critical clue that Bailey missed) or Bailey is aware that Campbell will better understand the significance of a particular aspect given his expertise. As a result, Bailey is more likely to be accurate in her reporting given Campbell's assistance, that is, in the dependent case. Importantly, we may know this even though *we* have not yet observed Campbell's report. Finally, even where Campbell is wrong, Bailey may still provide useful evidence, because the part of the judgment that is based on her own expertise is unaffected, so that even though she becomes less accurate overall, her report still has diagnostic value. In short, dependency *advantages* may exist under reasonable assumptions.

1.3. Present Research

Although this paper seeks to make the theoretical point that dependency can be advantageous, thus providing a corrective to previous claims in the literature, we also seek to determine how such advantageous instances are understood by lay reasoners. More precisely, we seek to determine whether lay reasoners a) endorse assumptions that entail such advantages, and b) reason appropriately in such cases (namely correctly identifying a dependency advantage, where one exists).

Assumption 1 - that the sending source Campbell is generally reliable - is provided directly to participants in the scenario description. However, both assumption 2 - that a recipient source is assisted (i.e., chance of error is reduced) by correct information from a sending source, and assumption 3 – that the degree of assistance outweighs the degree to which incorrect information misleads the recipient (i.e., increases error rates), are elicited directly from participants by asking for conditional probability estimates regarding the influence of the sending source (Campbell) on the recipient source (Bailey).

We expect that while participants may endorse these intuitive underlying assumptions, they will remain ignorant of the implied dependency advantages, instead showing a blanket preference for independence (e.g., Schum & Martin, 1982) or remain ignorant of the impact of (in)dependence altogether (e.g., Soll, 1999).

In Experiment 1 we test this in a baseline case where the two equally reliable sources either share information (dependent case), or remain independent. Experiment 2 then strengthens these findings via extended replication. We then build on this in Experiment 3, where the reliabilities of the two sources differ from one another. This third experiment is motivated not only by the desire to better map onto real-world cases (seldom are sources *exactly* equally reliable), but also to determine whether differences in reliability alter participants' assumptions and judgments regarding the nature of dependencies.

2. Experiment 1

Experiment 1 was designed to determine several interrelated points. Firstly, we sought to examine how well lay people's assumptions about the impact of a direct dependency between sources fit with the conditions outlined above. Do lay people understand, as our proof shows, that under certain information states (partial or contradicting), dependencies can provide an evidential advantage (relative to an independent equivalent)? Secondly, given their *own* stated assumptions, are lay people sensitive to the conditions under which a direct dependency is advantageous, or

do they adopt a blanket aversion to them? To explore this, we use a version of the plane crash scenario outlined above.

2.1. Method

Participants. Using Amazon Mechanical Turk, 200 US participants were recruited and participated online. Participants were native English speakers[5], with a median age of 33 years (*SD* = 11.06). 107 participants identified as female. Participants were paid $1.00 for their time (*Median* = 8.47 minutes, *SD* = 5.97).

Procedure & Design. Participants were presented with the plane crash outlined above (and illustrated in part in Fig. 2). Critically, participants were provided with a prior probability of the plane having crashed due to sabotage (P(Sabotage) = 0.5), along with reliability statistics for the two independent investigators, Bailey and Campbell (error rates – both false positive and false negative – of 20%[6]) when independent.

The procedure started with participants providing basic demographics (age, gender, location, and native language) before reading through the plane crash scenario (a full copy of these materials can be found in Supplementary Materials A) and providing conditional probabilities for their assumptions regarding the influence of a direct dependency on the reliability of a recipient source (Bailey). These conditional probability questions consisted of two 'if… then' statements, wherein participants needed to provide a probability (0-100) of Bailey making an error *given correct or erroneous information from Campbell*. Participants were provided with reminders that both Bailey and Campbell have the same 20% error rates when independent of each other when asked about the dependent case, specifically:

[5] 202 participants were originally recruited, but 2 were removed for not being based in the US nor being a native English speaker.

[6] I.e., $P(S_B|\neg Sabotage) = P(\neg S_B|Sabotage) = P(S_C|\neg Sabotage) = P(\neg S_C|Sabotage) = 0.2$

1. "If Bailey, before making her report, has seen Campbell's completed report - when that report is in fact CORRECT - what do you estimate is the probability of Bailey making a mistake now?"
2. "If Bailey, before making her report, has seen Campbell's completed report - when that report is in fact INCORRECT - what do you estimate is the probability of Bailey making a mistake now?"

These error rates for Bailey represent respectively (in the dependent case):

1. The conditional probabilities, $P(S_B=\text{Sabotage} \mid S_C=\text{Accident}, H=\text{Accident})$, that Bailey wrongly concludes sabotage despite being passed correct information from Campbell, and $P(S_B=\text{Accident} \mid S_C=\text{Sabotage}, H=\text{Sabotage})$, that Bailey wrongly concludes accident despite being passed correct information from Campbell; and
2. The conditional probabilities, $P(S_B=\text{Sabotage} \mid S_C=\text{Sabotage}, H=\text{Accident})$, that Bailey wrongly concludes sabotage when given erroneous information from Campbell, and $P(S_B=\text{Accident} \mid S_C=\text{Accident}, H=\text{Sabotage})$, that Bailey wrongly concludes accident when given erroneous information from Campbell.[7]

These conditional probabilities represent the key assumptions of the participant regarding the influence of the dependency, and for each participant are fed into individually-fitted BN models, such that optimal inferences could be determined for comparison, *given the participant's own background assumptions*.

Following the elicitation of conditional probabilities, participants saw the two comparison cases (dependent/independent) laid out above[8]. The specific wording of these were:

[7] These possibilities are symmetric in all experiments.
[8] Here we use the term "cases" to remain consistent in terminology throughout, however in the participant materials, "cases" were labelled "scenarios".

> Case 1: “You learn that Bailey, prior to completing her report, was accidentally given access to Campbell's completed report. As such, Bailey's report may be influenced by what Campbell has reported.”
>
> Case 2: “You learn that Bailey completed her report without ever seeing Campbell's completed report. As such, Bailey's report is not influenced by what Campbell has reported.”

Where case 1 is the dependent case, and case 2 the independent case.

Participants then compared these two cases using a qualitative comparison judgment:

> “Based on what you know at this point, which case (if either) provides more support for the plane having been sabotaged?” [“They are the same.” / “Case 1” / “Case 2”].

with the presentation order of those response options randomized across participants.

Participants then provided a confidence in that judgment:

> “How confident are you that your response is correct?” [Slider, 0 – 100%.].

And, finally, they provided probability estimates of the likelihood of sabotage in each case:

> “What is your current probability estimate of sabotage in each case, given what you know so far?” [Sliders from 0-100%.]

Crucially, these questions were asked across three time-points:

- t_0 *(Baseline)*: No reports from either investigator (only the background context – structure and parameters – have been provided).
- t_1 *(First report)*: Bailey gives a positive (sabotage) report.
- t_2 *(Second report)*: Campbell gives either a corroborating (sabotage) or contradicting (no sabotage) report.

In other words, t_0 assesses the extent to which participants agree with the interpretation of the scenario context (i.e., given a full understanding of the context, do they still agree that P(Sabotage) = 50%?), t_1 represents a state of partial information, and t_2 represents a state of complete corroborating or complete contradicting information.

Whether Campbell gave a corroborating or contradicting report was manipulated between-subjects.

2.2. Results

Bayesian statistics were employed throughout[9] using the JASP statistical software (JASP Team, 2018).

2.2.1. Conditional probabilities

As illustrated in Fig. 3, participants generally increased their estimates of Bailey's error rate when the secondary source Campbell is incorrect (median (red dashed line in Fig. 3) = 36%) – albeit with substantial variance (a likely marker of the differences in the way in which individuals determine background information); at the same time, they reduced their estimates of Bailey's error rate when provided with correct information from the secondary source (median (green dashed line in Fig. 3) = 15%).

To assess deviation from the starting (independent) reliability value of 20%, Bayesian t-tests were conducted on each conditional probability (in comparison to a test value of 20). Given the significant right-hand skew (evident in Fig. 3, and further evidenced by a Shapiro-Wilk p-values < .001), all estimates (and corresponding test value) were log transformed (x -> log(x +1) to account for 0 values, resulting in a test value of 1.322) before analysing. The analysis found

[9] Whilst Bayes Factors (BF_{10}: likelihood ratio of data given hypothesis, over data given null) > 3 may be considered substantial support for the alternative hypothesis, Bayes Factors < .33 are considered substantial support for the *null* (Jarosz & Wiley, 2014). For all analyses, an uninformative prior was used, unless otherwise specified. Wherever possible, sample sizes for a given analysis (*N*), and Bayesian Credibility Intervals (95% CI) are indicated.

that estimates of the impact of an "incorrect" secondary source showed decisive evidence for an increased estimation of error (N = 200; M = 1.518, 95% CI: [1.447, 1.558]), $BF_{10} = 8.06 * 10^{14}$. Conversely, the impact of a "correct" secondary source on error rates showed decisive evidence for a decrease[10] (N = 200; M = 1.194, 95% CI: [1.142, 1.247]), $BF_{10} = 3670.9$.

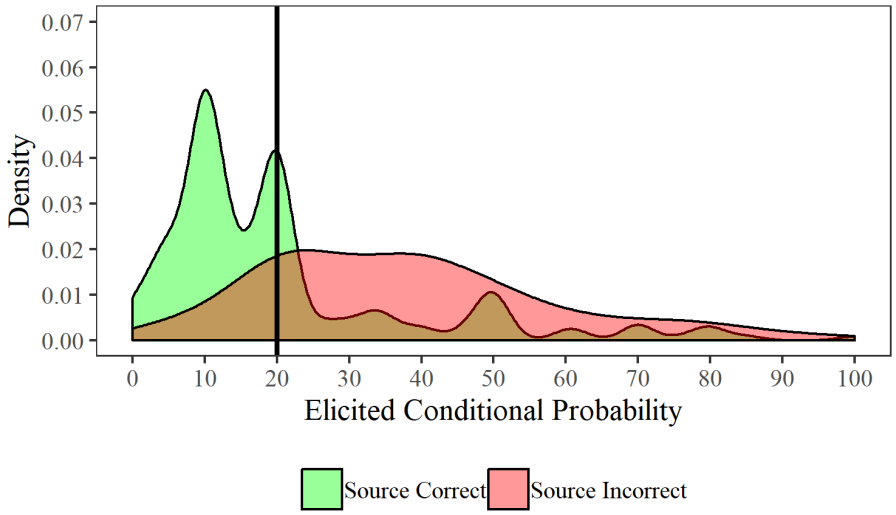

Figure 3. Density plots of the elicited conditional probabilities of the expected error rate for a dependent source (Bailey) with a standard (independent) error rate of 20% (vertical solid black line), when provided with correct (green) or incorrect (red) information from a second source (Campbell).

Using the gRain package in R (Højsgaard, 2012), the conditional probabilities elicited from each participant were used to fit the amended error rates for Bailey (as a recipient / dependent source) in a dependent case BN, creating individually fitted BNs for each participant (hereafter termed Behaviorally Informed Bayesian Networks; BIBNs). Each participant thus has a dependent and independent BIBN model, with the parameters drawn either from elicited conditional probabilities (Bailey in the dependent model, as explained above), or drawn from the parameters stated in the scenario description provided to participants (e.g., the prior probability of sabotage).

[10] It is worth noting that the large variance exhibited in the elicited conditional probabilities (and notably when a source receives incorrect information) may be reflective of participant interpretations of the impact of *background information* inherent to considerations of dependence (Schum 1994).

The posterior probabilities and qualitative judgments for each case (at each elicitation stage) generated from each BIBN model (representing each participant) were used in subsequent comparison analyses.

2.2.2. Qualitative judgments

To analyze participants' qualitative judgments of which of the two cases provided more support for the sabotage hypotheses (response options: dependent, independent, or "same"), a series of Bayesian contingency tables (to assess factors) were used to firstly compare participant judgments across elicitation stages (t_0(Baseline), t_1(First Report), and t_2(Second Report)), and conditions (contradicting vs corroborating second reports).

We then sought to assess whether or not participants' judgments deviated from the normative model (as in, e.g., Harris et al., 2014). Participant judgments were compared to the fitted BIBN model predictions on the group level across elicitation stages, before then also being compared on the individual level in terms of internal coherence, using Binomial tests for comparing correct responding to chance.

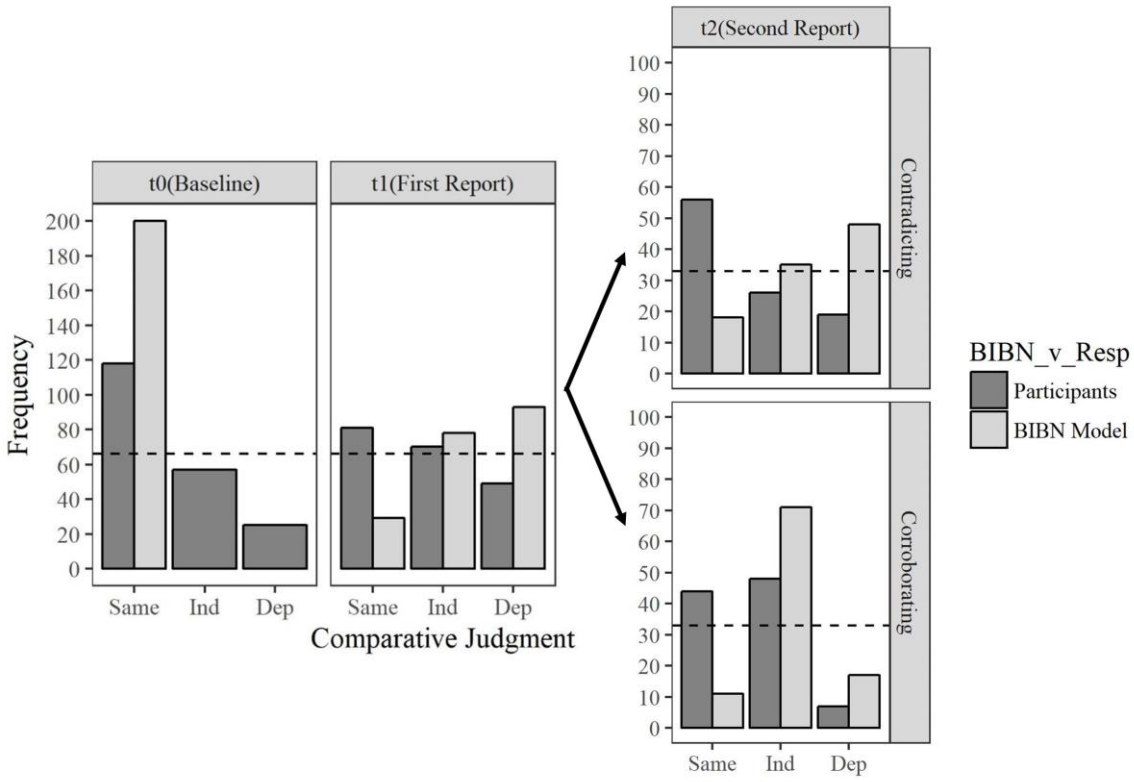

Figure 4. Frequency plots of qualitative comparison judgments, split by elicitation stage (t0, t1, t2). Corroborating and Contradicting second report conditions also shown. Dark bars represent participant responses, grey bars represent corresponding responses generated from the individually fitted Bayes net models (BIBN). Dashed line represents chance level (33%).

Participant Judgments. Turning first to participant judgments (dark grey bars, Fig. 4) across elicitation stages and conditions, a judgment (3) x elicitation stage (3) contingency table (N = 600), found substantial evidence for an effect of elicitation stage on participant judgments, BF_{10} = 5.72. In other words, participants' judgments of which case provided more support shifted as new evidence was presented. Further, when comparing the effect of condition on judgments, very strong evidence was found (N = 200), BF_{10} = 31.97, that participants were influenced by whether the second report corroborated or contradict the first: judgments of the independent case were more frequently taken to support the sabotage hypothesis. These results in effect serve to demonstrate a successful manipulation of evidence presentation.

Model Comparisons. We next compared the distribution of participants' judgments to those derived from the normative model. Specifically, we used the individually fitted BIBNs (light grey bars, Fig. 4) to generate the response distribution we should have seen, had participants' responded in line with their respective BIBNs. A series of Bayesian contingency tables were then used with a "data type" (Participant vs. BIBN prediction) factor. These analyses were conducted across the three elicitation stages (separating contradicting and corroborating conditions in the second report stage). In so doing, decisive evidence was found for the deviation of participant judgments from BIBN model predictions at baseline, t0 (N = 400), $BF_{10} = 3.59 * 10^{25}$, and first report, t1 (N = 400), $BF_{10} = 9.0 * 10^{6}$, stages. The former of these highlights that some participants did not consider the two cases (dependent or independent) to be the same prior to observing any evidence. This could reflect an interpretation that there is some form of pre-

emptive comparison of the cases, or could be an artifact of inattentive responding. The difference at the first report stage is, however, particularly notable; the trend of judgment preferences for participants ("Same", then "Independent", then lastly, "Dependent") is the mirror opposite of the judgments expected according to the BIBN models, where dependence is in fact the modal preference. This finding speaks to the failure of participants to understand the normative implications of their own assumptions when dealing with partial evidence (and their naïve undervaluing of dependent evidence as a consequence).

Finally, when turning to the comparison of participant judgments and model predictions at the second report stage, t2, both corroborating (N = 200), BF_{10} = 81188.76, and contradicting (N = 200), $BF_{10} = 1.27 * 10^6$, conditions reveal decisive evidence for group level proportions of judgments deviating from those predicted by participant BIBNs. In the corroborating condition, participants' majority response matched that of the BIBNs in viewing the independent case as stronger (albeit to an insufficient degree). In the contradicting condition, however, the rank order of participant preferences over response options "same/independent stronger/dependent stronger" incorrectly matched those of the partial information case (t_1), and, as there, they undervalued the strength of the dependent case (with it, once again, being the least preferred option), which in both cases should have been the most prevalent response according to the BIBNs.

Internal Coherence. To determine which participant judgments were erroneous (e.g., whether judgments of independence were more typically incorrect), the degree of internal coherence between participant judgments and BIBN model predictions was assessed on the individual level, with a "coherence" variable created for each participant judgment. This variable was either correct (1) – the participant judgment matches the model prediction, or else incorrect (0) and was then used to determine if participants made correct judgments more often than expected by chance (0.33) across elicitation stages and conditions, using Binomial tests.

At baseline, t0, correct responding (internal coherence) occurred decisively above chance levels (0.59, 95% CI: [0.52, 0.66]; $N = 200$), $BF_{10} = 1.68 * 10^{11}$. However, at the first report stage, *strong evidence for the null* (no difference from chance) was found (0.34, 95% CI: [0.28, 0.41]; $N = 200$), $BF_{10} = 0.088$, again reflective of the overlooking of the dominant dependent case in states of partial information. Similarly, substantial evidence was found for correct responding above chance level when the second report was corroborative (0.46, 95% CI: [0.36, 0.55]; $N = 100$), $BF_{10} = 3.39$, but when the second report was contradicting, correct responding showed *substantial evidence for the null* (no difference from chance; 0.36, 95% CI: [0.27, 0.45]; $N = 100$), $BF_{10} = 0.14$. This latter pair of results, again, indicate a failure to appreciate a dependency advantage (in this case when evidence is contradictory).

2.2.3. Confidence in Qualitative Judgments

The confidence in qualitative judgments was generally high across all judgments ($M = 67.49$, $SD = 23.84$). Using a Bayesian ANOVA ($N = 600$), confidence was shown to be unaffected by judgment, $BF_{Inclusion}$[11] = .49, elicitation stage, $BF_{Inclusion} = .014$ (substantial evidence for the null), or their interaction, $BF_{Inclusion} = .001$ (decisive evidence for the null). Confidence was, however, decisively higher when second reports ($N = 200$) were corroborative ($M = 74.4$, $SD = 23.61$) rather than contradicting ($M = 60.88$, $SD = 25.26$), $BF_{10} = 163.39$. This finding fits with the higher erroneous responding found in qualitative judgments when evidence is contradictory, and the more general argument that contradicting evidence is harder to integrate.

2.2.4. Probability estimates

We next turn to participant probability estimates. At each time step, participants estimated the probability of sabotage given the evidence presented so far – for both the dependent and independent cases. Using Bayesian repeated-measures ANOVA, participant estimates were first compared across elicitation stages and conditions, and then later compared to the estimates

[11] $BF_{Inclusion}$ is the change in odds from the sum of prior probabilities to the sum of posterior probabilities across models that include the effect.

derived from their BIBN models. Fig. 5 below shows the mean estimates for dependent (grey lines) and independent (black lines) case participant estimates (solid lines), BIBN estimates (dashed lines) across elicitation stages (within-facet), and conditions (facets).

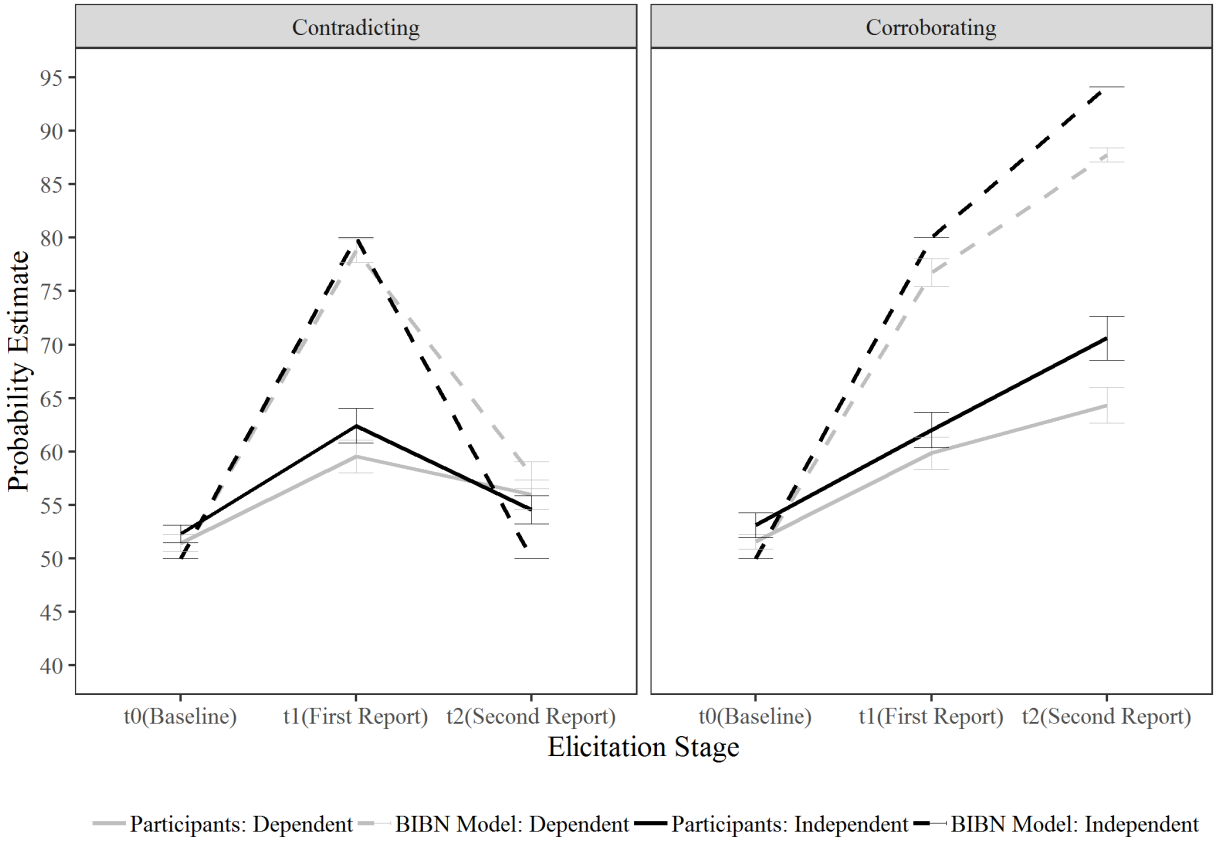

Figure 5. Mean participant estimates of the probability of sabotage across elicitation stages (t0, t1, t2), split by contradicting vs corroborating second report conditions. Dashed lines reflect BIBN model predictions, whilst solid lines reflect participant estimates. Grey lines illustrate probability estimates for the dependent case, and black lines probability estimates for the independent case. Error bars reflect standard error.

Participant Estimates. A Bayesian, repeated-measures ANOVA, including all relevant factors (within: case type, elicitation stage; between: condition) was run in a hierarchical model on participant probability estimates (N = 1200). The model including main effects for elicitation

stage, case type, and condition, as well as the interactions of elicitation stage x condition and case type x condition, enjoyed the strongest support, $BF_M = 56.25$, with decisive evidence overall, $BF_{10} = 2.62 * 10^{44}$. This model is hereafter referred to as $Model_P$. Looking at the effect terms, independent cases (black solid lines, Fig. 5) were considered to provide more support than dependent equivalents (grey solid lines, Fig. 5), $BF_{Inclusion} = 7.27$, substantiating qualitative judgment findings. There was also a main effect of elicitation stage (linearly increasing estimates across stages), $BF_{Inclusion} > 150$, indicating participants were generally sensitive to incoming information, including the impact of contradictory vs corroborating information, $BF_{Inclusion} > 150$ (decisive evidence was found for the elicitation stage x condition interaction in $Model_P$, along with a decisive main effect of condition, $BF_{Inclusion} > 150$).

Lastly, motivated by the evidenced case type x condition interaction in $Model_P$ ($BF_{Inclusion} = 8.08$), two separate Bayesian ANOVAs were conducted on estimates in the second report elicitation state, split by condition, testing the effect of case type in each condition ($N = 200$ in each condition). In the corroborative condition, participants assigned greater support to the independent case (relative to dependent), $BF_{10} = 223.23$, whilst in the contradicting condition there was substantial evidence for a null difference, $BF_{10} = 0.16$. This again fits the qualitative judgment data, wherein the corroborative case (which is typified by an independent case advantage) is easier for participants to determine (confidence was also higher).

Model Comparison. To address the question of how *sufficient* this updating is, participant data were compared to BIBN model predictions (dashed lines, Fig. 5). To appropriately explore model fit on the individual level, a Bayesian repeated measures ANOVA (as in Harris et al., 2014) was conducted on probability estimates with the additional inclusion of data type (Participant vs. BIBN prediction) as a within-subject factor ($N = 2400$).

Decisive evidence was found for the main effect of data type, $BF_{Inclusion} > 150$, indicating participant probability estimates were generally lower than the predictions of their BIBN models,

fitting with other findings of under-adjustment. Decisive evidence was also found for the interaction of data type and elicitation stage (indicating the insufficient updating increased over stages), $BF_{Inclusion} > 150$, condition (participant updating was more insufficient in the corroborating condition), $BF_{Inclusion} > 150$, and their 3-way interaction (the greater insufficiency of updating in the corroborating condition occurs in second report state), $BF_{Inclusion} > 150$. Data type did not interact with case (dependent or independent) both in isolation, $BF_{Inclusion} = 0.041$, or in conjunction with other factors. Accordingly, the model combining $Model_P$ with data type and the evidenced interactions above yielded the comparatively strongest fit, $BF_M = 914.89$, with decisive evidence overall, $BF_{10} = 1.49 * 10^{356}$. Taken together, this demonstrates that participant updating was insufficient relative to their BIBN model predictions (i.e., evidence is generally undervalued). Further, the interaction with condition (as with participant data alone) motivates the restricted analysis of second report estimates only, split by condition (but now also including data type).

In the corroborating condition, decisive evidence was found for the main effects of case type (independent > dependent), $BF_{Inclusion} = 898.02$, and data type (model > participant), $BF_{Inclusion} > 150$, but not their interaction, $BF_{Inclusion} = 0.90$. Thus, the model consisting of the two main effects only provided the comparatively strongest fit, $BF_M = 17.71$, with decisive evidence overall, $BF_{10} = 3.54 * 10^{41}$. This indicates that although participants generally undervalue the impact of a second, corroborative report, their assessment of the greater support provided in independent (relative to dependent) cases is broadly in line with model predictions. Conversely, in the contradicting condition, there was no evidence for a main effect of data type, $BF_{Inclusion} = 1.183$, and strong evidence for the null for both the main effect of case type, $BF_{Inclusion} = 0.08$, and their interaction, $BF_{Inclusion} = 0.039$. Consequently, the model consisting solely of data type was the comparatively strongest fit, $BF_M = 5.26$, but there was no evidence for this model overall, $BF_{10} = 1.75$. Although this appears to indicate a reasonable fit between participant data and model

predictions for contradicting evidence cases, such a null difference may in fact be attributable to under-valuing evidence *twice*, rather than accurate updating.[12]

2.3. Discussion

We find that lay people generally endorse the assumption that correct information passed to a recipient source from a secondary source (of equal reliability) will *decrease* the likelihood of error in the recipient, as determined by the elicited conditional probabilities. We find also that there is substantial variance in participant estimations of the degree to which a recipient source will be misled by incorrect information; this may be an indicator of the different context-based assumptions participants are making (i.e., different perceptions of the influence of background information).

Critically, we show that the majority of participants *should*, according to their own BIBN model predictions, also judge dependencies as advantageous (in terms of the degree of support provided to the hypothesis) in both partial and contradicting information states. However, we find that participants fail to appreciate this implication in their qualitative judgments, instead eschewing preferences for dependence (which is only appropriate in complete-corroborating information states). We find confidence in these judgments (irrespective of accuracy) to generally be high, with the exception that contradicting cases are marked by lower confidence relative to corroborating equivalents – an indicator of the increased difficulty of the former.

Finally, we find that participant probability estimates fit with previous findings on under-adjustment in the face of new evidence (e.g., Phillips & Edwards, 1966). Importantly, participant probability estimates in partial and contradicting information states reveal a preference to assign more support (wrongly) to the independent case, substantiating the qualitative judgment findings.

[12] A Bayesian repeated measures ANOVA of first to second report estimates in the contradicting condition reveals an interaction of data type with elicitation stage from first report to second report, $BF_{Inclusion} = 1.95 * 10^{15}$, highlighting the differential in updating between data and model prediction.

3. Experiment 2

We further assessed the robustness of these findings in Experiment 2.

First, to test the generalisability of elicited conditional probabilities, as well as subsequent judgment and probability estimate findings of Experiment 1, the primary hypothesis under investigation, which was previously always whether the crash was due to sabotage, was now counterbalanced between-subjects as either the crash was due to an accident, or the crash was due to sabotage. In conjunction with this, the conditional probabilities elicited were expanded from two to four questions to reflect all possible states of the world. Previously, these were only concerned with changes in Bailey's chance of error when provided with correct/incorrect information from Campbell, but now this was separated also out by hypothesis (i.e., whether Campbell was correct/incorrect about the crash being due to sabotage / an accident). In this way it was possible to detect potential hypothesis-specific trends.

Second, the order or reports was manipulated between subjects, such that at the partial information stage (one report back) half the participants saw a report from Bailey (as in Experiment 1), and the other half saw Campbell's report. This allowed us to distinguish between patterns of judgments and estimates caused by the presentation of partial information *in general*, and patterns of judgments made as a consequence of partial information pertaining to the recipient source *specifically*. To elucidate, it allows us to determine whether previously observed difficulties in dealing with partial information are a result of that partial information coming from a potentially dependent source, or are more general. Further, we could also detect whether the sequence of reports (e.g., Bailey first or second) affected judgments and estimates at t2 (i.e., whether errors at t2 were a consequence of which source reported at t1).

Whilst the changes regarding the hypothesis of interest (accident vs sabotage) are not predicted to have any influence on results (both in terms of counterbalancing and conditional probabilities for the two hypotheses being the same), the order of reports is expected to have

specific impact at the first report (partial information) stage. More precisely, when presented with Bailey first, then results should echo those of Experiment 1 (participants failing to appreciate instances of dependency advantages), but when presented with Campbell first, there *should* be no difference between independent and dependent cases. This is expected to be more intuitive to participants, as no uncertain inference across the dependency is required (i.e., the value of Campbell's report is independent of whether or not Bailey has seen it).

3.1. Method

Participants. Participants were recruited using the same protocol as in Experiment 1. A sample size of 200 was predetermined, in line with Experiment 1. Of the 203 participants recruited, 3 were removed for either not having English as a native language, and/or being based outside of the US. Of the 200 remaining participants, 124 identified as female. The median age was 30 years (*SD* = 10.59). Participants were paid $1.00 for their time (*Median* = 8.94 minutes, *SD* = 6.59).

Procedure & Design. The procedure and design of Experiment 2 followed that of Experiment 1, with the following exceptions:

First, the conditional probability questions presented to participants at the beginning of the task are now separated by hypothesis. These four (rather than two) questions are as follows:

1. "The plane did in fact crash due to **sabotage**, and Campbell **CORRECTLY reports** the crash as being caused by sabotage. Bailey has seen Campbell's report. What do *you* estimate is the probability of Bailey then **INCORRECTLY reporting that the plane crashed due to an accident**? (*Remember: Without seeing Campbell's report, it would be 20%*)"
2. "The plane did in fact crash due to **an accident**, and Campbell **CORRECTLY reports** the crash as being caused by an accident. Bailey has seen Campbell's report. What do *you* estimate is the probability of Bailey then **INCORRECTLY reporting that**

the plane crashed due to sabotage? (*Remember: Without seeing Campbell's report, it would be 20%*)"

3. "The plane did in fact crash due to **sabotage**, and Campbell **INCORRECTLY reports** the crash as being caused by an accident. Bailey has seen Campbell's report. What do *you* estimate is the probability of Bailey then **INCORRECTLY reporting that the plane crashed due to an accident**? (*Remember: Without seeing Campbell's report, it would be 20%*)"
4. "The plane did in fact crash due to **an accident**, and Campbell **INCORRECTLY reports** the crash as being caused by sabotage. Bailey has seen Campbell's report. What do *you* estimate is the probability of Bailey then **INCORRECTLY reporting that the plane crashed due to sabotage**? (*Remember: Without seeing Campbell's report, it would be 20%*)"

Second, whether the first report confirmed sabotage or accident was counterbalanced (hereafter referred to as the target hypothesis condition), and the qualitative comparison judgement and probability estimate questions matched this by asking about support provided for / probability of sabotage or accident respectively. Thus, in the accident target hypothesis condition, ACCIDENT was replaced by SABOTAGE.

For qualitative judgments, participants in this condition saw the question:

> "Based on what you know *at this point*, **which case (if either) provides more support for the plane having crashed due to ACCIDENT**?".

For probability estimates, the same participants saw the questions:

> "What is your **current probability estimate** of the plane having crashed because of ACCIDENT in each case, given what you know so far? *Click the bars below, indicating*

between 0% (left-most point) and 100% (right-most point).

Probability of crash due to SABOTAGE in **Case 2** (%) [Slider 0 -100%]

Probability of crash due to SABOTAGE in **Case 1** (%) [Slider 0 -100%]"

Third, which investigator provided the first report was also manipulated between-subjects (hereafter referred to as reporter order condition), such that half the participants received a report from Bailey first (as in Experiment 1), and the other half received a report from Campbell first.

As in Experiment 1, whether the second report agreed (corroboration) or disagreed (contradiction) with the first was also manipulated between subjects.

3.2. Results

The analytical procedure follows that of Experiment 1, with the further addition of the between-subjects factors; target hypothesis, and reporter order. The former is not expected to influence results, and as such given the absence of a main effect, will be removed as a factor from all subsequent analyses.

3.2.1. Conditional probabilities

Following the same protocol as Experiment 1, and further evidenced by Fig. 6 and Shapiro-Wilk p-values < 0.001, to reduce right-hand skew, all estimates were log transformed prior to analysis. We then tested whether the target hypothesis condition affected the degree to which participants' estimates of the conditional probabilities for source accuracy (i.e., given dependent reports) differed from the experimenter provided rates for the individual, independent sources, using a repeated-measures ANOVA (correct/incorrect x sabotage/accident). Whilst a decisive evidence for a main effect was found for the influence of correct vs incorrect information, $BF_{Inclusion} = 7.755 * 10^{39}$, strong evidence for the null was found for the influence of hypothesis, $BF_{Inclusion} = 0.076$. Consequently, the model with only a main effect of correct vs incorrect information was the best fit, $BF_M = 46.66$, and decisive overall, $BF_{10} = 7.785 * 10^{39}$.

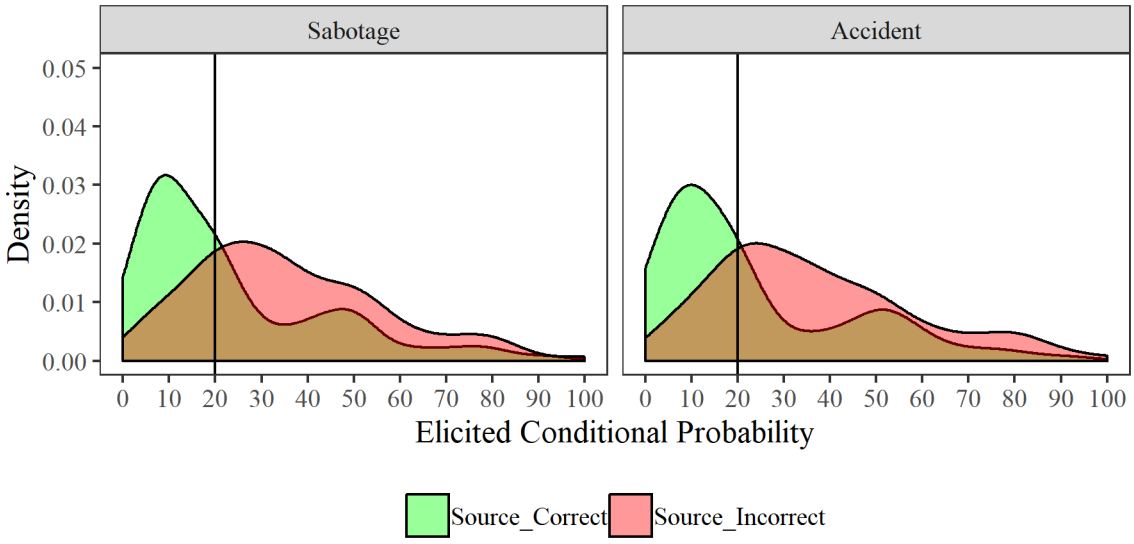

Figure 6. Density plots of the elicited conditional probabilities of the expected error rate for a dependent source (Bailey) with a standard (independent) error rate of 20% (vertical solid black line), when provided with correct (green) or incorrect (red) information from a second source (Campbell), split by target hypothesis condition.

A series of Bayesian t-tests were then used to determine whether participants' estimates of source accuracy given a second (corroborating or conflicting report) differed from the 20% error in the independent case (log transformed to a value of 1.322), so as to determine the degree to which participants estimated correct or incorrect information to assist or mislead a recipient. First, very strong evidence is found for a difference (expected decrease in error rates) when correct information is provided in either the sabotage (N = 200, M = 1.216, 95% CI: [1.159, 1.272]), BF_{10} = 52.82, or accident hypotheses, (N = 200, M = 1.2, 95% CI: [1.14, 1.261]), BF_{10} = 132.18. Conversely, decisive evidence is found for the expected increase in error rates given incorrect information in both sabotage (N = 200, M = 1.487, 95% CI: [1.447, 1.527]), $BF_{10} = 1.313 * 10^{11}$, and accident ($N$ = 200, M = 1.495, 95% CI: [1.455, 1.535]), $BF_{10} = 9.847 * 10^{11}$, hypotheses. This suggests, in replication of Experiment 1, that participants generally considered recipients to

be compromised by incorrect information and assisted by correct information – irrespective of hypothesis.

3.2.2. Qualitative judgments

Following the analysis protocol of Experiment 1, participant's qualitative judgments were first analyzed using contingency tables to assess the influence of elicitation stages and conditions (contradicting/corroborating second reports, reporter order, and target hypothesis counterbalancing). Following this, the analyses assess how participant judgments compared to those expected by BIBN models across elicitation stages. As in Experiment 1, this was first assessed at the group level (proportions of judgments, using contingency tables), followed by the individual level (internal coherence, using Binomial tests).

Participant Judgments. Accordingly, using a series of Bayesian contingency tables (N = 600), strong evidence was found for the effect of elicitation stage on participant judgments, BF_{10} = 11.93, and substantial evidence for the effect of reporter order condition, BF_{10} = 3.88. As expected, there was substantial evidence for a null effect of target hypothesis counterbalancing, BF_{10} = 0.145, and as such this factor is removed from subsequent analyses. Although there was also strong evidence for a null effect of second report condition, BF_{10} = 0.051, this may be attributable to the condition only occurring at the final elicitation stage.

Illustrated below in Fig. 7 are the qualitative judgments of participants (dark grey bars), and their BIBN model predictions (light grey bars) based on the elicited conditional probabilities of those same participants.

In line with Experiment 1, we find that participants generally eschew judgments of dependence being superior (dark grey bars, Fig. 7). Although this is appropriate (vis-à-vis BIBN model predictions; light grey bars, Fig. 7) in cases where dependence is either inferior (corroborating reports) or irrelevant (Campbell is the sole reporter; First Report, Campbell First facet), this is naïve in the face of cases where dependencies may yield an informational advantage (Bailey is

the sole reporter; First Report, Bailey First facet; and when reports contradict). These discrepancies are borne out in the group-level contingency table analyses of participant vs model predictions across elicitation stages.

Model Comparisons. In fact, decisive evidence was found for participant judgments differing from those predicted by their BIBN models at baseline (a substantial number of participants erroneously not selecting "same"; $N = 400$), $BF_{10} = 2.197 * 10^{31}$. Substantial evidence was found for this deviation when Bailey reports first (insufficient number of participant judgments for dependent case advantage, in line with Experiment 1; $N = 202$), $BF_{10} = 4.964$, and decisive evidence was found for this deviation when Campbell reports first (insufficient appreciation for the independent and dependent cases being the same; $N = 198$), $BF_{10} = 4.553 * 10^{8}$. As there was strong evidence for a null effect of reporter order on participant judgments at the second report stage ($N = 200$), $BF_{10} = 0.062$, these conditions were collapsed to investigate corroborating and contradicting second reports. In the former, corroboration led to strong evidence for there being too few independent case advantage judgments among participants ($N = 194$), $BF_{10} = 24.15$, and in the latter, there was decisive evidence for contradiction leading to too few judgments of both dependence and independence ($N = 206$), $BF_{10} = 6.649 * 10^{9}$.

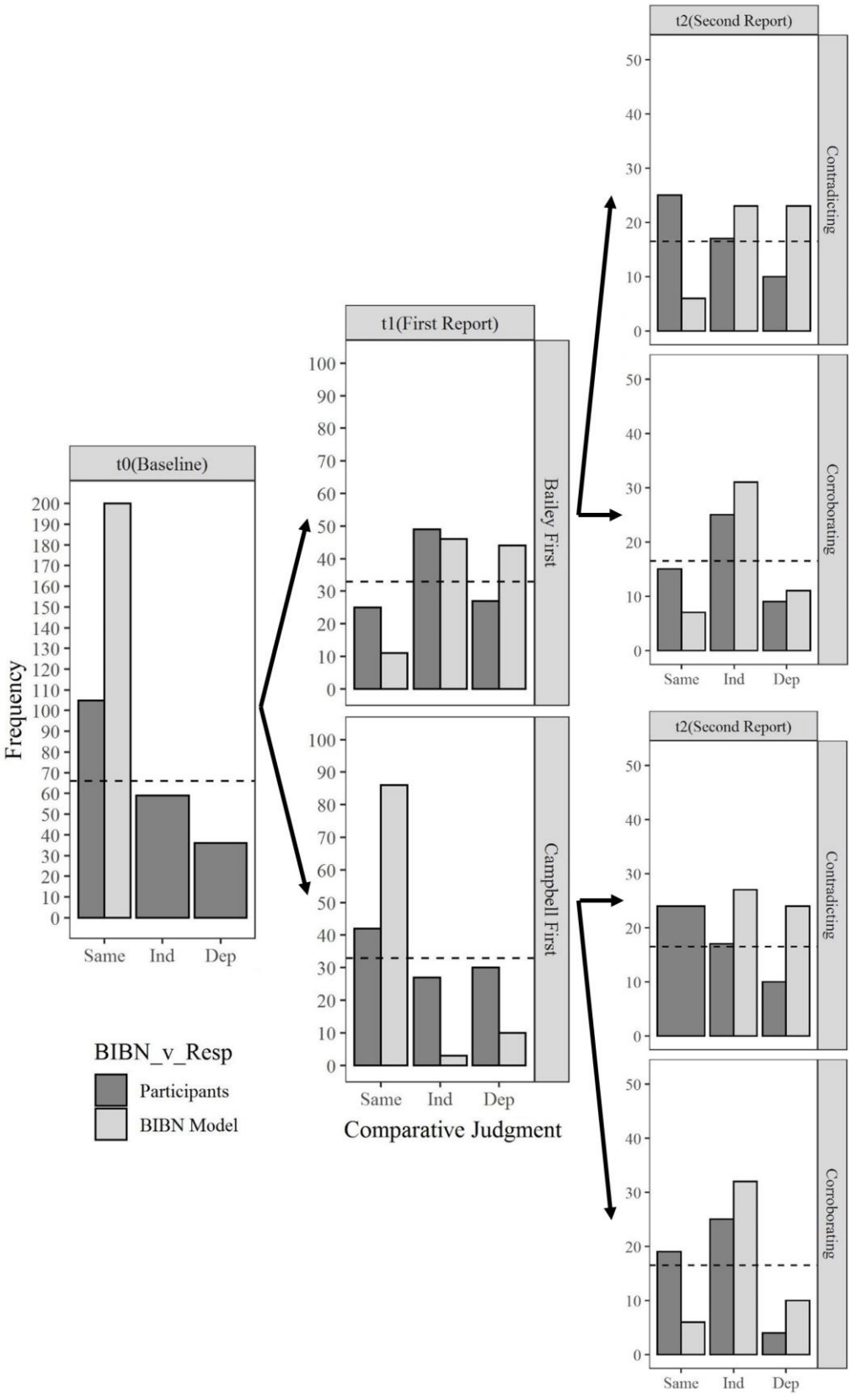

Figure 7. Frequency plots of qualitative comparison judgments, split by elicitation stage (t0, t1, t2), reporter order (from t1 onwards; middle column), and corroborating vs contradicting second report conditions (right-hand column). Dark bars represent participant responses, grey bars represent

corresponding responses generated from the individually fitted Bayes net models (BIBN). Dashed line represents chance level (33%).

Internal Coherence. To confirm these deviations (and to be in line with the analysis protocol of Experiment 1), participant vs model comparisons were assessed on the individual level using Binomial tests. More precisely, if participant judgment and the BIBN judgment for that participant agreed, then the judgment was marked as correct (1), but if they failed to agree, were marked as incorrect (0). This variable could then be compared to chance (.33) for performance comparison. Consequently, decisive evidence was found for performance at baseline being greater than chance level (0.525, 95% CI: [0.456, 0.593]; $N = 200$), $BF_{10} = 845{,}142.3$. Then when Bailey reported first, substantial evidence for the null was found (0.376, 95% CI: [0.288, 0.474]; $N = 101$), $BF_{10} = 0.193$, whilst no evidence was found when Campbell reported first (0.414, 95% CI: [0.322, 0.513]; $N = 99$), $BF_{10} = 0.568$. Similarly, no evidence was found for performance differing from chance when reports corroborated (0.412, 95% CI: [0.319, 0.512]; $N = 97$), $BF_{10} = 0.523$, whilst substantial evidence for the null was found when reports contradicted (0.282, 95% CI: [0.204, 0.375]; $N = 103$), $BF_{10} = 0.194$. In sum, although dependency advantages are overlooked in specific partial and contradicting information states, they appear to be part of a general trend in inaccuracy when considering cases of direct dependence.

3.2.3. Confidence in qualitative judgments

As in Experiment 1, confidence in qualitative judgments was generally high across all judgments ($M = 67.69$, $SD = 24.36$). An initial Bayesian ANOVA found substantial evidence for a null effect of target hypothesis counterbalancing condition (N = 600), $BF_{10} = 0.239$, as expected, and was thus excluded from further analysis. In a second Bayesian ANOVA ($N = 600$), confidence was also again shown to be unaffected by judgment, $BF_{Inclusion} = 0.262$, elicitation stage, $BF_{Inclusion} = 0.011$, reporter order condition, $BF_{Inclusion} = 0.229$, and second report condition, $BF_{Inclusion} = 0.086$,

with substantial to very strong evidence for the null for these main effects, and their interaction terms. Also, as in Experiment 1, there was very strong evidence for confidence being higher when second reports (N = 200) were corroborative (M = 75.26, SD = 21.75) as opposed to contradicting (M = 62.47, SD = 26.08), BF_{10} = 96.58. This again speaks to the higher error rates in judgments when evidence is contradictory, as well as to the idea that contradictory evidence is generally more difficult to integrate.

3.2.4. Probability estimates

As in Experiment 1, participant probability estimates of the likelihood of sabotage in the two different cases (dependent and independent; grey vs black lines, Fig. 9) across elicitation stages (baseline, first report, and second report; within-facet, Fig. 9) were assessed using a Bayesian repeated measures ANOVA. Initially, this model included only the target hypothesis counterbalancing condition to check for unexpected influence on probability estimates. Having found strong evidence for no influence of target hypothesis counterbalancing (N = 1200), $BF_{Inclusion}$ = 0.029, this factor was excluded from subsequent analyses. Consequently, a first model that included the two within-subject factors (case and elicitation stage), as well as the two between-subject factors (second report condition; facet rows, Fig. 9, and reporter order; facet columns, Fig. 9). In a follow up analysis, participant BIBN model predictions were included as a further within-subject factor.

Participant Estimates. The first analysis, focusing on participant estimates alone (solid lines, Fig. 9; N = 1200), revealed decisive evidence for main effects of elicitation stage (increasing trend from baseline to second report), $BF_{Inclusion} > 150$, and second report condition (corroborating > contradicting), $BF_{Inclusion} = 2.891 * 10^{14}$, whilst there was no effect of reporter order condition, $BF_{Inclusion}$ = 0.572, and substantial evidence for anull effect of case type (dependent = independent), $BF_{Inclusion}$ = 0.303. Consequently, there was only decisive evidence for the interaction of elicitation stage and second report condition (the increasing trend across stages is

more substantial in the corroborating condition), $BF_{Inclusion} = 2.405 * 10^{13}$, and the strong evidence for the three-way interaction of elicitation stage, second report condition, and reporter order condition (the difference in trends between corroborating and contradicting conditions is more pronounced when Campbell has reported first), $BF_{Inclusion} = 17.33$. The model containing the requisite terms for this three-way interaction, along with case type and its interaction with reporter order, enjoyed the strongest fit, $BF_M = 85.65$, was decisive overall, $BF_{10} = 4.928 * 10^{25}$, and is hereafter termed $Model_{P2}$. Taken together, we find participants are sensitive to introduced evidence, both via sequential presentation, and of valence (positive or negative support provided), including the degree to which these factors may be influenced by reporter order. However, these estimates do not differ on the basis of case type (independent / dependent).

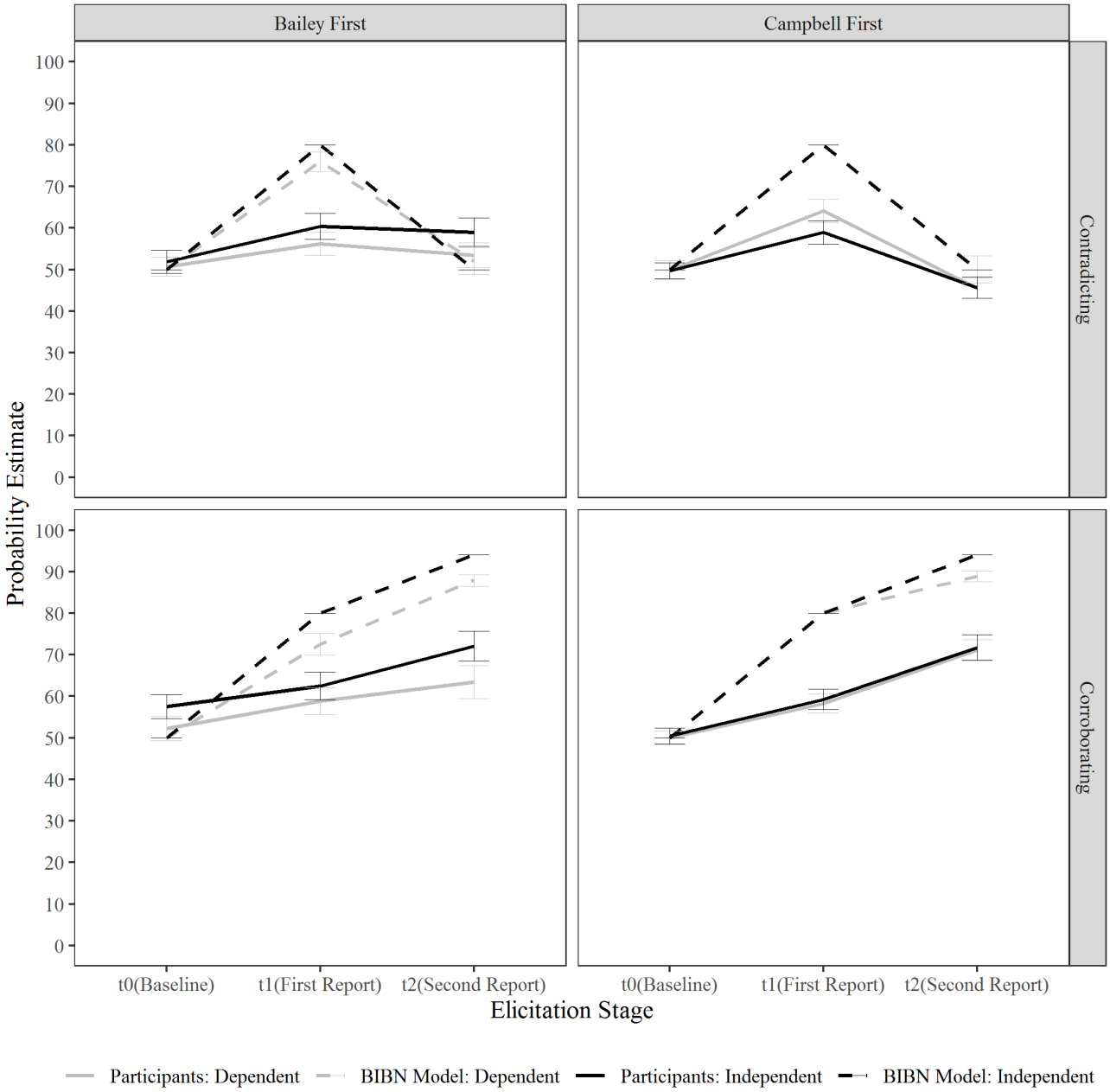

Figure 8. Mean participant estimates of the probability of sabotage across elicitation stages (t0, t1, t2), split by contradicting vs corroborating second report conditions (rows) and reporter order (Bailey first vs Campbell first) conditions (columns). Dashed lines reflect BIBN model predictions, whilst solid lines reflect participant estimates. Grey lines illustrate probability estimates for the dependent case, and black lines probability estimates for the independent case. Error bars reflect standard error.

Model Comparisons. To assess the degree to which participant estimates deviated from normative expectation, the BIBN model predictions for each participant were added to the above

analysis protocol (dashed lines, Fig. 9; included as an additional factor; data type). However, given the size of the resultant ANOVA, the analyses are split by reporter order condition (columns, Fig. 9), with results focusing on the effect terms involving the BIBN model comparison. Irrespective of reporter order condition, we find the same trend of results. More precisely, this analysis found decisive evidence for a main effect of data type (model predictions > participant estimates) when Bailey reports first (N = 1212), $BF_{Inclusion} = 3.423 * 10^{11}$, and when Campbell reports first (N = 1188), $BF_{Inclusion} = 4.306 * 10^{36}$. This deviation (wherein participants underestimate the value of evidence) increased over elicitation stages, (Bailey first reporter) $BF_{Inclusion} = 1.431 * 10^{14}$, and (Campbell first reporter) $BF_{Inclusion} = 3.012 * 10^{23}$. These underestimation trends were more pronounced in the corroborating conditions, (Bailey first reporter) $BF_{Inclusion} = 8.846 * 10^{11}$, and (Campbell first reporter) $BF_{Inclusion} = 1588.91$.

Although $Model_{P2}$ illustrated that participants were sensitive to the sequence and direction of evidence, the above shows that this sensitivity is generally insufficient. This is further evidenced by the strongly to decisively evidenced interactions of data type and second report condition in both the Bailey first reporter, $BF_{Inclusion} = 53.31$, and Campbell first reporter, $BF_{Inclusion} = 474.21$, conditions. However, as case type (dependent vs independent; grey vs black line, Fig. 9) was not found to interact with data type, this suggests that the lack of differentiation between independent and dependent cases among participants at the aggregated (group) level is not predicted by their aggregated BIBN model predictions.

The models that only included all the above terms (and their requisite base-terms) enjoyed the strongest fit, (Bailey first reporter) $BF_M = 255.71$, and (Campbell first reporter) $BF_M = 1728.56$, and were decisive overall, (Bailey first reporter) $BF_{10} = 4.108 * 10^{105}$, and (Campbell first reporter) $BF_{10} = 2.795 * 10^{199}$.

3.3. Discussion

The purpose of Experiment 2 was to test the robustness of the effects found in Experiment 1 by manipulating the target hypothesis (sabotage vs accident), and checking for whether this influenced the elicited conditional probabilities and subsequent judgments and estimates. Further, we tested participant sensitivity to the order of reporters (Bailey first, vs Campbell first), and replicated Experiment 1 findings.

Importantly, we find no influence of target hypothesis across all dependent variables, and further show that conditional probabilities (i.e., the general endorsement of assumptions regarding the influence of dependencies) are not influenced by target hypothesis either (see Fig. 6), speaking to the validity of Experiment 1 findings.

Turning first to qualitative judgments, we find participants are sensitive to the order of reporters, with more participants correctly appreciating that independent and dependent cases are equivalent when only a sending source (Campbell) has reported. Furthermore, we replicate the general findings of Experiment 1, wherein participants fail to appreciate the dependency advantages dictated by their own (elicited) assumptions when only a receiving source has reported (Bailey), or the two reports contradict. We again note that participant confidence in their judgments (irrespective of accuracy) remains high throughout the task.

Participant probability estimates reveal that although participants are sensitive to the sequence and valence of evidence, their updates are generally insufficient relative to their fitted (BIBN) model predictions. We additionally note that when looking at the group level data for probability estimates (Fig. 8), differences in independent versus dependent case estimates are not readily apparent (a trend that fits with the difficulties in determining dependency advantages in the qualitative judgment data).

4. Experiment 3

In Experiments 1 and 2, we looked at a somewhat idealised scenario in which the two sources (when independent) were exactly equal in their reliability. Given the integral nature of the assumption (also held by participants) that the reliability of a source is altered when a direct dependency is present, a logical follow-on is to explore the assumptions (and subsequent potential dependency advantages) when sources are unequal in their reliability. For instance, if we consider a higher and lower reliability pair of sources (e.g., a senior vs junior doctor), then there is an intuitive difference in the assumed impact of a directional dependency between them. More precisely, information of a diagnosis from a senior doctor, passed on to a junior is more likely to reduce the latter's probability of making a mistake, whilst the reverse (junior to senior) is arguably less probable.

These intuitions also expose the incorporation of further background information (e.g., the perception of the reliability of the senior by the junior, and vice versa). As a consequence, we again return to the method of extracting the conditional probabilities on the individual level to create the Bayesian comparison.

In line with these intuitions we predict that when a recipient source is lower in reliability than a sending source, participants will make the same assumptions as in Experiments 1 and 2 (i.e., the chance of error in the recipient is reduced by correct information and increased by incorrect). However, when the recipient is higher in reliability than the sender, we expect there to be muted influence of the sender on the recipient's error rates (i.e. the senior doctor is neither assisted nor misled by notes from the junior doctor, when they can also both assess the patient).

4.1. Method

Participants. Participants were recruited using the same protocol as in Experiment 1. A sample size of 200 was predetermined, in line with Experiment 1. Of the 201 participants recruited (50

per group, see below), 102 were female. The median age was 32 years (*SD* = 11.46). Participants were paid $1.00 for their time (*Median* = 8.54 minutes, *SD* = 5.51).

Procedure & Design. The procedure and design of Experiment 3 followed that of Experiment 1, with the following exception:

In the present experiment, only one of either Bailey or Campbell was stipulated to be as reliable as in Experiments 1 and 2 (20% error rates), with the other was stipulated as being higher in reliability (5% error rates). Whether the recipient (Bailey) or the sender (Campbell) was the higher reliability of the two was manipulated between-subjects [Recipient High Reliability; RHR / Sender High Reliability; SHR]. This difference was described as follows (recipient high reliability condition shown, low in braces):

"Secondly, the analysis is notoriously difficult, and Bailey and Campbell differ in their reliability:

- Bailey has a 5% [20%] probability of mistakenly indicating that sabotage has occurred (and a 5% [20%] probability of mistakenly indicating sabotage has not occurred).
- Campbell has a 20% [5%] probability of mistakenly indicating that sabotage has occurred (and a 20% [5%] probability of mistakenly indicating sabotage has not occurred)."

This change led to one further amendment to the procedure, wherein the elicitation of conditional also included a further reminder of the recipient reliability:

1. "If Bailey, before making her report, has seen Campbell's completed report - **when that report is in fact CORRECT** - what do you estimate is the probability of Bailey making a mistake now? (Remember: Without seeing Campbell's report, it would be 5% [20%])"
2. "If Bailey, before making her report, has seen Campbell's completed report - **when that report is in fact INCORRECT** - what do you estimate is the probability of Bailey making a mistake now? (Remember: Without seeing Campbell's report, it would be 5% [20%])"

4.2. Results

The analytical procedure follows that of Experiment 1, with the further addition of a between-subjects factor; reliability difference.

4.2.1. Conditional probabilities

Using a series of Bayesian t-tests, participants' estimates of the conditional probabilities were compared relative to the report recipient's stipulated rate for the independent case (5% in the RHR condition, and 20% in the SHR condition) to determine the degree to which participants estimated correct or incorrect information from a high/low reliability sender to assist or mislead a low/high reliability recipient. Once more, given the right-hand skew evident in Fig. 9, and further evidenced by Shapiro-Wilk p-values < 0.001, in accordance with Experiments 1 and 2, all data and test values were log transformed (x -> log(x + 1), to combat 0 values) prior to analysis.

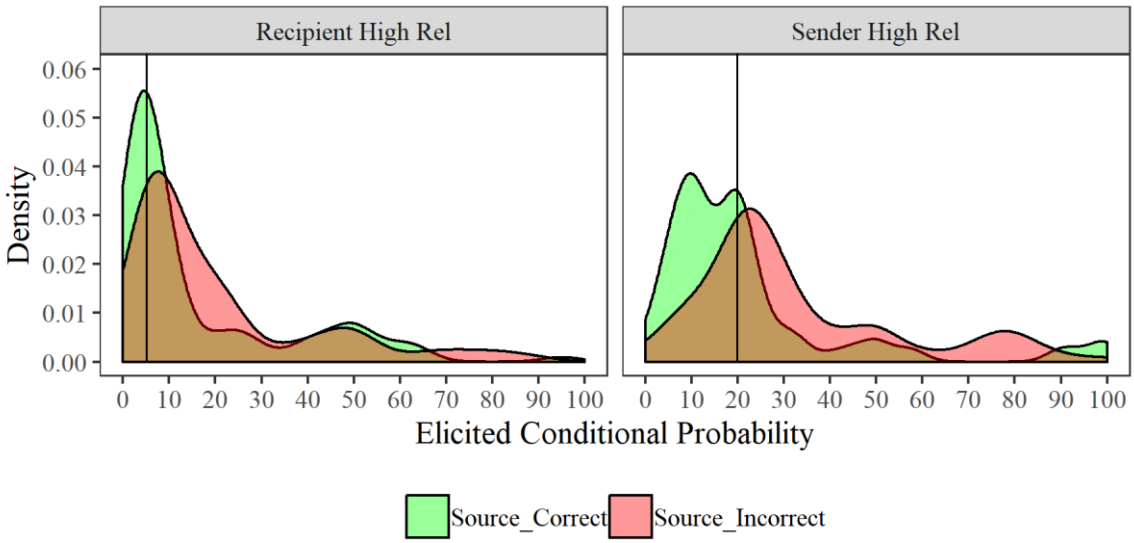

Figure 9. Density plots of the elicited conditional probabilities of the expected error rate for a dependent source (Bailey), when provided with correct (green) or incorrect (red) information from a second source (Campbell). Solid black lines reflect independent *recipient* (Bailey) source error rates, such that an in the recipient high reliability condition (left-hand facet) the independent recipient error rate is 5% (and the sender has an error rate of 20%), whilst in the sender high reliability condition (right-hand facet) the independent recipient error rate is 20% (and the sender has an error rate of 5%)..

Turning first to the RHR condition (left-hand facet of Fig. 9), relative to the 5% starting (independent) error rate of the recipient (log value = 0.778), when the sender provided incorrect information (red distribution), recipient error rates were judged to increase (N = 101, M = 1.165, 95% CI: [1.092, 1.237]), BF_{10} = 9.341 * 10^{14}. However, when the sender provided correct information (green distribution), participants *still* judged the recipient error rates to increase (N = 101, M = 0.974, 95% CI: [0.884, 1.064]), BF_{10} = 465.18. This suggests participants considered high reliability recipients to *always* be compromised by lower reliability sources, irrespective of the veracity of the information provided by the latter, however, this influence appears to be somewhat muted (with the distribution peaks remaining around the original 5% independent error rates).

In the converse condition, when the sender is high in reliability (SHR; and the recipient lower in reliability; right-hand facet of Fig. 9), elicited conditional probabilities are instead compared to the 20% (independent) starting error rate of the recipient (log value = 1.322). In this condition, when the sender passes on incorrect information (red distribution), recipient error rates are again judged to increase (N = 101, M = 1.425, 95% CI: [1.36, 1.49]), BF_{10} = 11.10. However, when the sender passes on correct information (green distribution), substantial evidence is found for a decrease in recipient error rates, in line with Experiments 1 and 2 (N = 101, M = 1.214, 95% CI: [1.14, 1.29]), BF_{10} = 4.99. Consequently, when extracted for use in the individual model fits (BIBN process, see Experiment 1), it is expected that this latter condition should yield the same overall model prediction of a dependency advantage in the partial and contradicting information states.

Thus, we may conclude that participants generally endorse assumption 2 (and 3). More precisely, participants consider receiving sources to be more accurate when provided with correct information from a sending source. However, this is only endorsed when the sender is more

accurate than the receiver (the sender being more likely to be accurate generally, and the receiver having more capacity to be improved by correct information).

4.2.2. Qualitative judgments

Following the analysis protocol of Experiment 1, participants qualitative judgments were first analyzed using contingency tables to assess the influence of elicitation stages and conditions (both contradicting / corroborating, and recipient/sender high reliability). Following this, the analyses are split by reliability condition (given the different conditional probabilities underpinning them) to assess how participant judgments compared to those expected by BIBN models across elicitation stages. As in Experiment 1, this was first assessed at the group level (proportions of judgments, using contingency tables), followed by the individual level (internal coherence, using Binomial tests).

Participant Judgments. Accordingly, using a series of Bayesian contingency tables (N = 603), very strong evidence for the null was found for the effect on participant judgments of elicitation stage, BF_{10} = 0.028, or reliability condition, BF_{10} = 0.064. However, whether the second report contradicted or corroborated the first (N = 201) was again found to have a substantial effect on participant judgments (more preferences for the independent scenario in the corroboration condition), BF_{10} = 3.29. Although the former of these findings depart from those found in Experiment 1, this may be attributable to the collapsing across reliability conditions, and not unexpected given the higher variance entailed.

Sender High Reliability. Fig. 10 below illustrates the qualitative judgments of participants (dark grey bars), and their BIBN model predictions (light grey bars) based on the elicited conditional probabilities of those same participants. In line with Experiments 1 and 2, both the partial (first report) and contradicting information states show a modal preference in the BIBN model predictions for the dependent case, whilst this is not the case in the participant judgments

themselves. These discrepancies are borne out in the group-level contingency table analyses of participant vs model predictions across elicitation stages.

Model Comparisons. In fact, participant judgments were found to decisively differ from those predicted by their BIBN models at baseline (a substantial number of participants erroneously not selecting "same"; $N = 200$), $BF_{10} = 9.61 * 10^{17}$, and first report (insufficient number of participant judgments for "dependent" case advantage, in line with Experiments 1 and 2; $N = 200$), $BF_{10} = 889.80$, stages. Strong evidence was also found for this deviation at the corroborating second report stage (too few "independent" case advantage judgments among participants; $N = 100$), $BF_{10} = 19.59$, and very strongly differ at contradicting second report stage (too few "dependent" case advantage judgments among participants, in line with Experiments 1 and 2; $N = 100$), $BF_{10} = 81.84$. Taken together, these suggest participants not only failed to adequately understand the dependency advantage implications when information was partial or contradictory, but more generally had difficulty with comparing independent and dependent cases when reliabilities differ between sources.

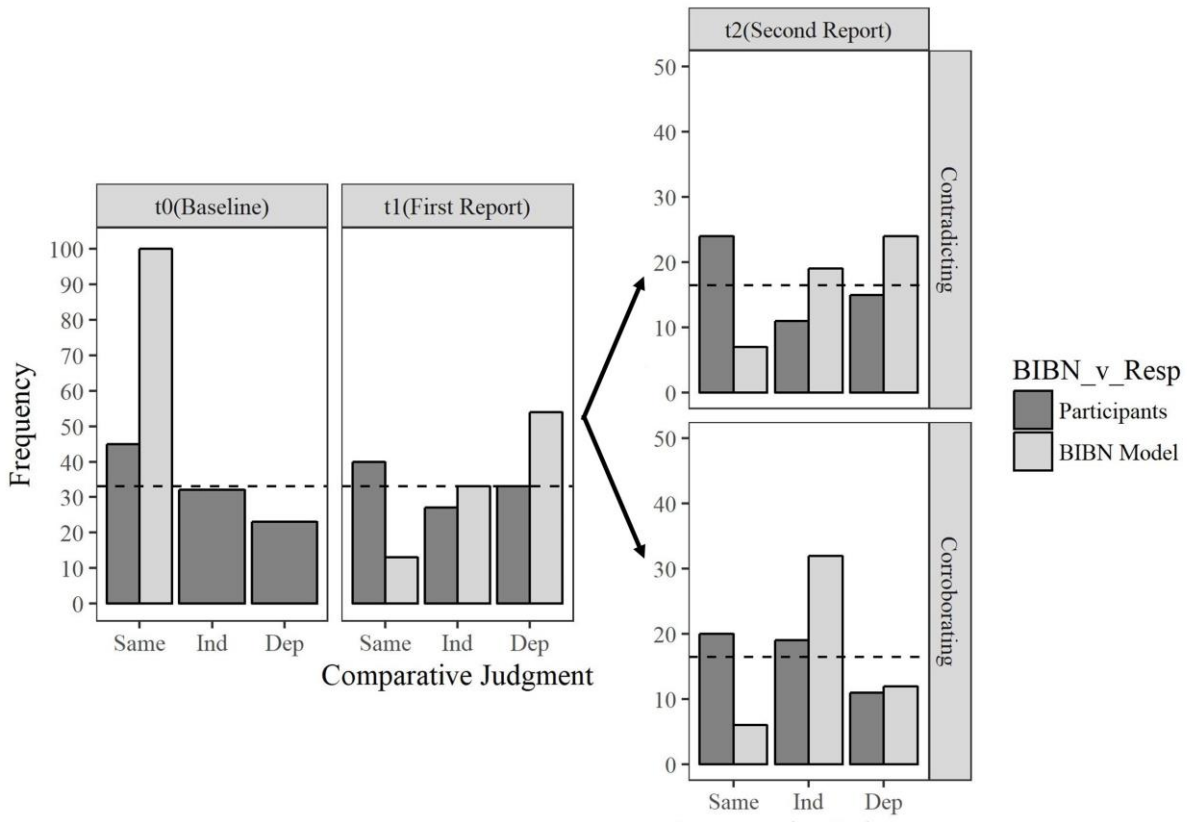

Figure 10. Frequency plots of qualitative comparison judgments in the Sender High Reliability condition, split by elicitation stage (t0, t1, t2) and (at t2) corroborating vs contradicting second report condition. Dark bars represent participant responses, grey bars represent corresponding responses generated from the individually fitted Bayes net models (BIBN). Dashed line represents chance level (33%).

Internal Coherence. To confirm this supposition (and to be in line with the analysis protocol of Experiment 1), participant vs model comparisons were assessed on the individual level using Binomial tests. More precisely, if participant judgment and the BIBN judgment for that participant agreed, then the judgment was marked as correct (1), but if they failed to agree, were marked as incorrect (0). This variable could then be compared to chance (.33) for performance comparison. Consequently, no evidence was found for performance at baseline differing from chance level (0.45, 95% CI: [0.36, 0.55]; $N = 100$), $BF_{10} = 2.75$, and substantial evidence for the null was found at the partial (first report) stage (0.34, 95% CI: [0.26, 0.44]; $N = 100$), $BF_{10} = 0.12$, second report corroborating stage (0.34, 95% CI: [0.22, 0.48]; $N = 50$), $BF_{10} = 0.17$, and second report contradicting stage (0.36, 95% CI: [0.24, 0.50]; $N = 50$), $BF_{10} = 0.19$. In sum, although dependency advantages are overlooked in the partial and contradicting cases, they appear to be part of a general trend in inaccuracy, possibly stemming from the added complexity of a difference in reliability between sources.

Recipient High Reliability. As can be seen in Fig. 11, when the recipient source is higher in reliability, the conditional probabilities elicited from participants result in BIBN model predictions that favor independence across corroborating and partial (first report) and contradicting information states (light grey bars in central and right-hand facets). This is unsurprising given that the sufficient assumptions (namely that recipient sources are assisted by correct information to a greater or equal degree than they are misled by incorrect information) are not met in this condition.

Model Comparisons. The potential discrepancies between participant judgments and BIBN model predictions at the group level (i.e., the overall frequency of judgments) were again assessed using Bayesian contingency tables. Consequently, we once again find participant judgments decisively differ from BIBN predictions at baseline (insufficient "same" judgments among participants; N = 202), BF_{10} = 5.57 * 10^{12}, strongly differ at first report (insufficient "independent" case advantage judgments among participants; N = 202), BF_{10} = 11.12, corroborating second report (insufficient "independent" case advantage judgments among participants; N = 102), BF_{10} = 21.24, and decisively at contradicting second report (again, insufficient "independent" case advantage judgments among participants; N = 100), BF_{10} = 1357.38, stages. Taken together, this again speaks to the explanation that as information states entail greater complexity when making an inference about the value of a dependency (i.e. partial and contradictory information states) erroneous responding increases, as although the modal participant judgments match those expected by BIBN models in baseline ("same"), first report ("independent"), and second report corroborating stages ("independent"), there is still a significant degree of error (despite independence dominance) – particularly notable in the more complex contradictory second report condition (where participant modal preferences for "same" are misplaced).

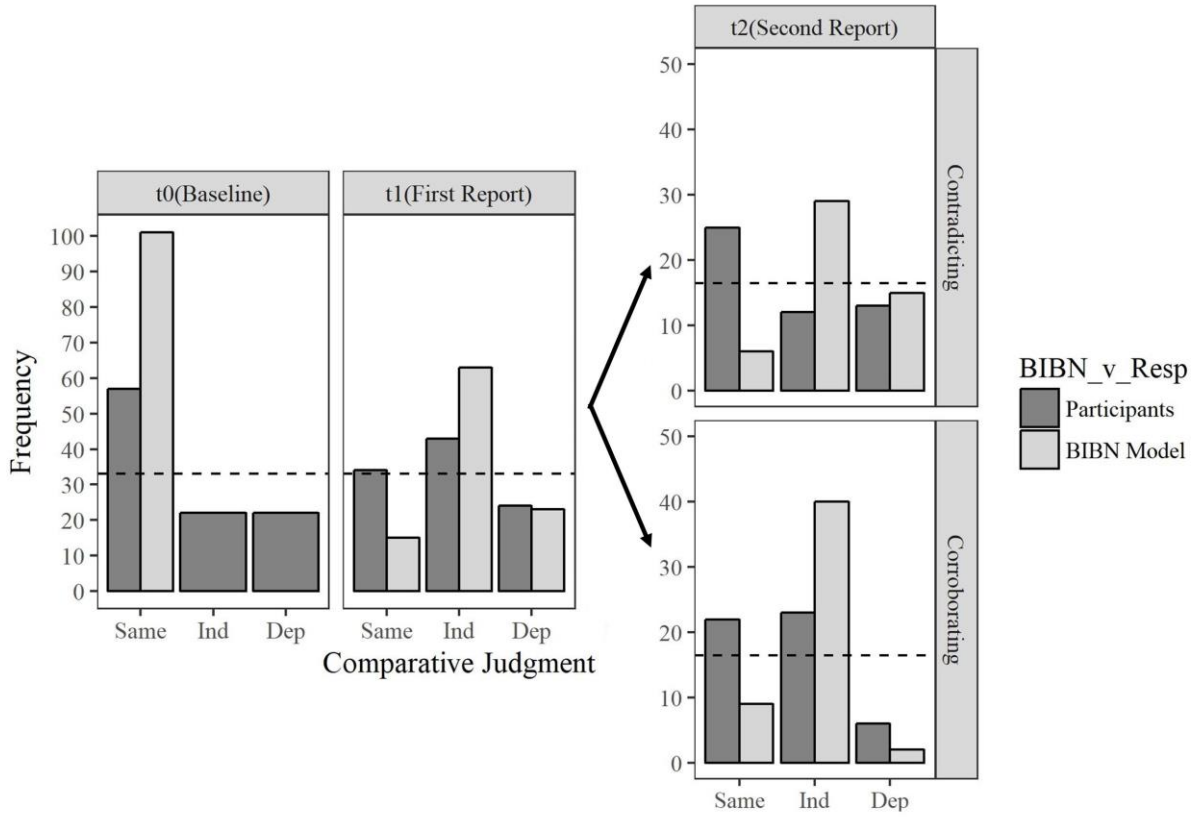

Figure 11. Frequency plots of qualitative comparison judgments in the Recipient High Reliability condition, split by elicitation stage (t0, t1, t2) and (at t2) corroborating vs contradicting second report condition. Dark bars represent participant responses, grey bars represent corresponding responses generated from the individually fitted Bayes net models (BIBN). Dashed line represents chance level (33%).

Internal Coherence. As in the sender high reliability condition, participant judgments were assessed relative to BIBN model predictions on the individual level (via the creation of a correct (match) / incorrect (mismatch) variable, to then be compared to chance level (0.33) using Bayesian Binomial tests), across elicitation stages. Accordingly, correct responding was found to be decisively greater than chance level at baseline (0.56, 95% CI: [0.47, 0.66]; $N = 101$), $BF_{10} = 14022.71$, and strongly at the first report stage (0.49, 95% CI: [0.39, 0.58]; $N = 101$), $BF_{10} = 22.12$. However, at the second report stage no evidence was found for correct responding being greater than chance level in the corroborating condition (0.47, 95% CI: [0.34, 0.61]; $N = 51$),

BF_{10} = 1.50, nor in the contradicting condition (0.34, 95% CI: [0.22, 0.48]; *N* = 50), BF_{10} = 0.17 (substantial evidence for the null). Taken together, responses were more accurate in the partial information state (first report stage) than in the sender high reliability condition (where dependence was dominant), but otherwise remained equivalently inaccurate. This latter aspect of the results further suggests that the inequality in reliabilities prevented accurate determination of which case provided support.

4.2.3. Confidence in qualitative judgments

As in Experiments 1 and 2, confidence in qualitative judgments was generally high across all judgments (*M* = 65.36, *SD* = 26.32). Using a Bayesian ANOVA (*N* = 603), confidence was also again shown to be unaffected by judgment, $BF_{Inclusion}$ = 0.003, elicitation stage, $BF_{Inclusion}$ = 0.01, and reliability condition, $BF_{Inclusion}$ = 0.013, with decisive to very strong evidence for the null for these main effects, and their interaction terms. Also, as previously, substantial evidence was found for confidence again being higher when second reports (*N* = 201) were corroborative (*M* = 72.56, *SD* = 26.89) rather than contradicting (*M* = 62.45, *SD* = 26.24), BF_{10} = 4.47. This again speaks to the higher error rates in judgments when evidence is contradictory, as well as the more general argument that contradictory evidence is more difficult to integrate.

4.2.4. Probability estimates

As in Experiments 1 and 2, participant probability estimates of the likelihood of sabotage in the two different cases (dependent and independent; grey vs black lines, Fig. 12) across elicitation stages (baseline, first report, and second report; within-facet, Fig. 12) were assessed using a Bayesian repeated measures ANOVA. Initially, a hierarchical model was deployed that included the above two within-subject factors (case and elicitation stage), as well as the two between-subject factors (second report condition; facet rows, Fig. 12, and reliability condition; facet columns, Fig. 12). In a follow up analysis, participant BIBN model predictions were included as a further within-subject factor.

Participant Estimates. A Bayesian, repeated-measures ANOVA, including all relevant factors (within: case type, elicitation stage; between: reliability and second report conditions) was run on participant probability estimates (solid lines, Fig. 12; N = 1206). This analysis revealed decisive evidence for main main effects of elicitation stage (increasing trend from baseline to second report), $BF_{Inclusion}$ = 1.156 * 10^{15}, and second report condition (corroborating > contradicting), $BF_{Inclusion}$ = 5.196 * 10^{11}, whilst there was no evidence for an effect of reliability condition, $BF_{Inclusion}$ = 0.58, and very strong evidence for a null effect of case type (dependent = independent), $BF_{Inclusion}$ = 0.011. Additionally, there was decisive evidence for the interaction of elicitation stage and second report condition (the increasing trend across stages is more substantial in the corroborating condition), $BF_{Inclusion}$ = 5.196 * 10^{11}, and strong evidence for a three-way interaction of elicitation stage, second report condition, and reliability condition (the difference in trends between corroborating and contradicting conditions is more pronounced when the sender is higher in reliability), $BF_{Inclusion}$ = 11.102.

Thus, the model containing only the requisite terms to include this three-way interaction (i.e., precluding case type and its derivate interactions) enjoyed the strongest fit, BF_M = 205.47, was decisive overall, BF_{10} = 2.676 * 10^{36}, and is hereafter termed $Model_{P3}$. Taken together, we find participants are sensitive to introduced evidence, both via sequential presentation, and of valence (positive or negative support provided), including the degree to which these factors may be influenced by strength (i.e., the reliability of reporters). However, these estimates do not differ on the basis of case type (independent / dependent).

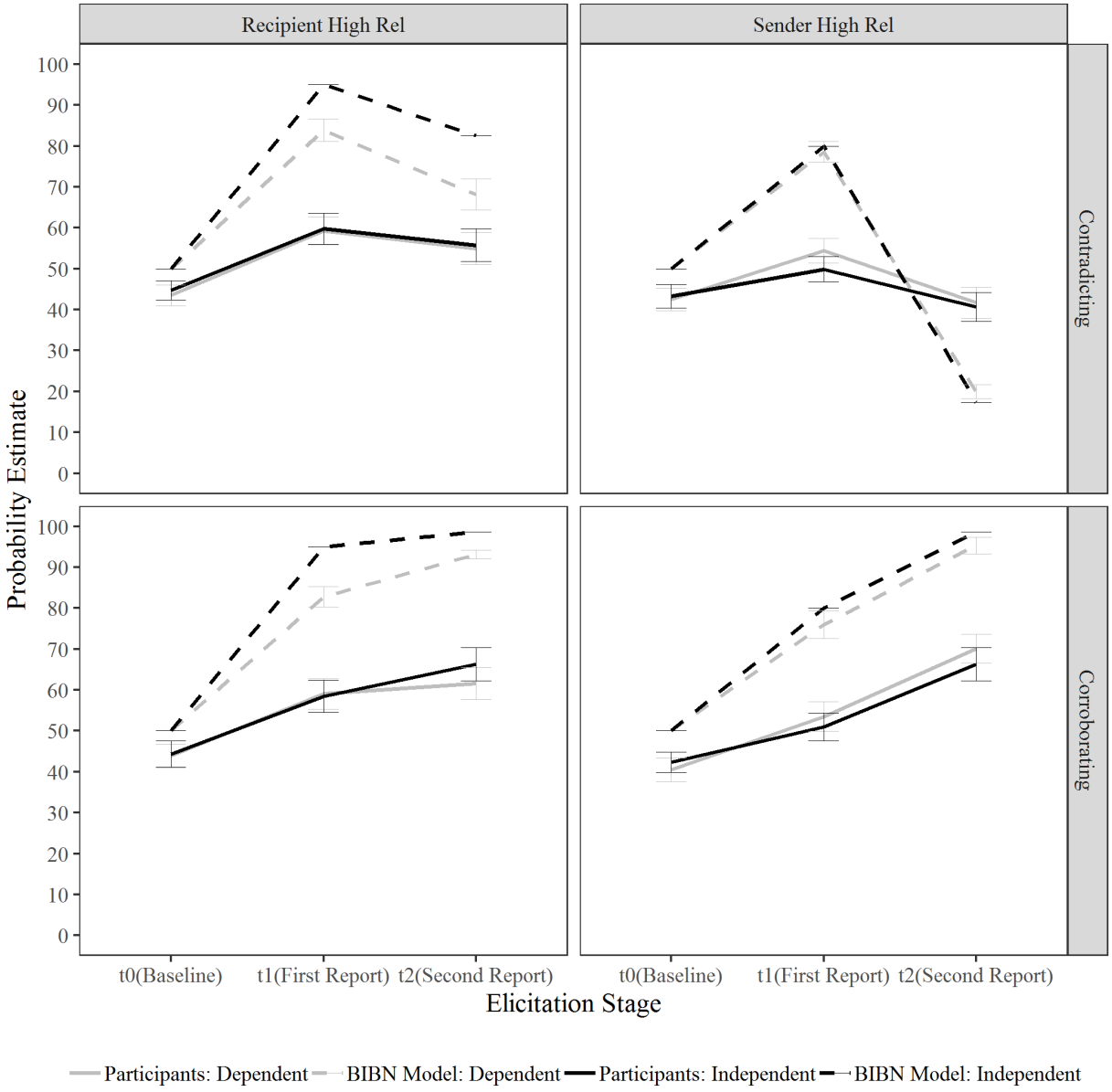

Figure 12. Mean participant estimates of the probability of sabotage across elicitation stages (t0, t1, t2), split by contradicting vs corroborating second report conditions (rows) and reliability (Recipient vs Sender high) conditions (columns). Dashed lines reflect BIBN model predictions, whilst solid lines reflect participant estimates. Grey lines illustrate probability estimates for the dependent case, and black lines probability estimates for the independent case. Error bars reflect standard error.

Model Comparisons. To assess the degree to which participant estimates deviated from normative expectation, the BIBN model predictions for each participant were added to the above analysis protocol (dashed lines, Fig. 12; included as an additional factor; data type; N = 2412),

with results focusing on the effect terms involving this comparison. This analysis found decisive evidence for a main effect of data type was found (model predictions > participant estimates), $BF_{Inclusion} = 5.07 * 10^{13}$. Decisive evidence was found for this deviation (wherein participants underestimate the value of evidence) increasing over elicitation stages, $BF_{Inclusion} = 8.11 * 10^{14}$, a trend further exacerbated by participant failures to sufficiently appreciate the influence of reliability (participants' underestimation is more pronounced when reporters are higher in reliability), $BF_{Inclusion} = 6.25 * 10^{15}$, and the second reporter condition (underestimation trends are further pronounced in the corroborating condition), $BF_{Inclusion} = 6.25 * 10^{15}$. This insensitivity is further exposed by decisive evidence for the four-way interaction of data type, elicitation stage, reliability and second report conditions, $BF_{Inclusion} = 3.07 * 10^{9}$, wherein contradiction from higher reliability sources (top-right facet, Fig. 12) should entail substantial downward revision, and thus participants are overestimating, whilst a lower reliability contradicting source (top-left facet, Fig. 12) does not negate the prior positive report, and thus participant conservative updates result in underestimation.

Although $Model_{P3}$ illustrates that participants are sensitive to evidence factors (sequence, valence, and strength), the above shows this sensitivity is generally insufficient. This is further shown by the decisive evidence for the interaction of data type and second report condition (high reliability contradictors should have a more substantial refuting effect than lower reliability contradictors, and corroboration is more impactful from a higher reliability source, but participants fail to adequately capture this), $BF_{Inclusion} = 8.11 * 10^{14}$. Similarly, the decisive evidence for the interaction of data type and reliability condition, $BF_{Inclusion} = 8.11 * 10^{14}$, reflects a similar increased deviation from normative expectation when the sender is reliable (right-hand column, Fig. 9). Lastly, the decisive evidence for the three-way interaction of data type, second report and reliability conditions, $BF_{Inclusion} = 8.11 * 10^{14}$, reflects the additive nature of these two condition-based deviation differences.

Lastly, substantial evidence was found for the interaction of case type (dependent vs independent; grey vs black line, Fig. 12) with data type, $BF_{Inclusion}$ = 4.57, indicating an insensitivity in participants to account for the superiority of the independent case (relative to dependent) – most clearly exemplified by the differences between black and grey dashed lines (BIBN models) and the (lack of) differences between black and grey solid lines (participants) in the left-hand columns of Fig. 12. This fits with the modal "same" preference in participant qualitative judgments – notably in first report and second report stages. Relatedly, the substantial evidence for the interaction of case and reliability condition, $BF_{Inclusion}$ = 6.14, is again reflective of the independent case superiority when the recipient is high in reliability – an effect borne out in qualitative judgment BIBN model predictions.

The model that only included all the above terms (and their requisite base-terms) enjoyed the strongest fit, BF_M = 15786.45, and was decisive overall, $BF_{10} = 1.52 * 10^{376}$.

4.3. Discussion

The introduction of reliability differences between the two sources in Experiment 3 yielded further insights into the way in which dependencies are (dis)advantageous. For instance, when a sender is higher in reliability than a recipient, participants generally endorse (via elicited conditional probability estimates) the same assumptions as in Experiment 1 and 2, wherein the (lower reliability) recipient is considered to benefit from correct information when it is passed to them by the (higher reliability) sender. As a consequence, we replicate the judgment and probability estimate findings of Experiments 1 and 2 (only this time with differences in reliability, rather than equal source reliabilities), in that, individual BIBN model predictions point to a dependency *advantage* in cases of partial and contradicting information states, but this advantage is still overlooked in participant responses.

Conversely, we find that in instances where a recipient source is considered to be higher in reliability than a sender, participants do not consider such a source to benefit from correct

information – given that this information originates from a lower reliability source. Instead, the impact of a secondary source is considered to have a net negative influence on the recipient (i.e., it can only be a detriment to the high reliability recipient's accuracy, relative to an independent equivalent). Consequently, BIBN model judgments and probability estimates then reflect this assumption in a global preference for independence across corroborating, partial and contradicting information states.

We find, however, that in participant qualitative judgments (but also corroborated by probability estimate data) there is a failure to appreciate this superiority of the independent case – most clearly highlighted in the contradicting information state. A majority of participants instead consider the cases (dependent and independent) the same.

This finding suggests two important possibilities. First, the introduction of differing reliabilities exposes a potential blanket or heuristic preference for reliability cues over structural differences. In this way, less attention is paid to the possible implication of the introduction of a direct dependency between the two sources, and thus the two cases are genuinely considered to be similar. Second, that high levels of erroneous responding – particularly in partial and contradicting information states – are not necessarily tied to the special case of dependency advantages, but rather may be more reflective of participants being overwhelmed by the complexity of these comparisons. This may be particularly acute when having to consider differences in the reliability of sources, and may in turn explain a "shallow, reliability cue" strategy outlined above.

5. General Discussion

The consideration of dependencies is critical to reasoning accurately about evidence. Dependence comes in many forms, such as a sharing of evidence (Schum, 1994) or background (Bovens & Hartmann, 2003; Madsen, Hahn, & Pilditch, 2018) between sources, or on an aggregate level, the degree to which reports are correlated (Einhorn, Hogarth, & Klempner, 1977;

Hogarth, 1978; Berg, 1993; Ladha, 1995). The standard conception of dependencies has been to consider their introduction as a form of redundancy (and thus inferior to an independent equivalent) in terms of the support provided to a hypothesis. Using a novel approach in which we disentangle structural dependence from observation (e.g., observed, correlated reports), we make the novel theoretical point that there exist – *under reasonable assumptions* – dependency *advantages* when observations are either a) partial, or b) contradict one another across a dependency relation. We provide a proof for the existence of cases in which these advantages hold, and we further demonstrate empirically that many lay reasoners endorse the outlined sufficient assumptions. However, we also show that despite endorsing such assumptions, lay reasoners struggle, both qualitatively and quantitatively, to understand the dependency advantage implications, instead preferring to assume that dependence is inferior.

Critically, by exploring the impact of dependencies when sources are equally reliable (Experiments 1 and 2), and when reliabilities are unequal (Experiment 3), we reveal a more nuanced picture of the difficulties of considering reliability, structural dependence, and crucially, differences in information states.

First, we note that sufficient assumptions for dependency advantages are endorsed as a function of the reliability of the sources involved. More precisely, when sources are either considered equally reliable (Experiment 1 and 2), or the sending source is considered higher in reliability than the recipient (Sender High Reliability condition; Experiment 3), then lay reasoners generally endorse the assumptions that recipients may *benefit* from this sharing of information (assumption 2; elicited via conditional probabilities) between generally reliable sources (assumption 1). As a consequence, in partial and contradicting states of information, a dependency advantage should result. However, when a recipient source is considered more reliable than the sender (Recipient High Reliability condition; Experiment 3), then reasoners do not consider the recipient to gain any meaningful benefit – and thus independence should

dominate across all information states. This interaction of evidence structure and reliability parameters (and the resultant variance in assumptive estimations) points to the difficulties intrinsic to considerations of dependencies. However, we note that our formal approach - Bayesian networks that disentangle structure from observation, fitted to participant estimations of the conditional influence of a dependency – highlights a fruitful avenue for navigating this complexity, extracting tailored normative predictions for meaningful comparisons on both the individual and group level.

Second, this disentanglement reveals that reasoners not only fail to appreciate dependency *advantages* in partial and contradicting information states, but such failures may reflect a subset of *general difficulties* in dealing with such information states. This is illustrated by the failure of participants to prefer *independence* when it is entailed by their own assumptions across partial and contradicting information states (Recipient High Reliability condition; Experiment 3), instead judging there to be no difference between dependent and independent cases. We note the potential common psychological difficulty in such cases is that partial and contradicting information states are more computationally taxing, requiring the integration across multiple possible routes of explanation (i.e., one has to accurately integrate both the possibility that incorrect or correct information – unknown to you – has been passed, *and* its possible influence on a receiver), whilst corroborative cases (and sender only partial information states, see Experiment 2) may rely on shallower, presence vs absence reasoning. We also note that dependency advantage cases (requiring partial or contradicting information) are themselves harder to infer from real world "observations" alone, without the supplemental understanding of the causal structure inherent to that context.

The present work has both theoretical and applied ramifications. Our approach and accompanying proof, along with the empirical demonstration of the underlying assumptions in action, make the novel theoretical point that there exist systematic structural conditions under

which dependencies are evidentially advantageous. We note that empirical work on individual versus collective judgment has shown communication among crowd members to lead to both decrements (Lorenz, Rauhut, Schweitzer, & Helbing, 2011) and improvements (Jönsson et al., 2015; Becker, Brackbill, & Centola, 2017) on group accuracy. Formal frameworks such as Condorcet's (1785) Jury Theorem or the Diversity Prediction Theorem (Page, 2008) help understand these seeming contradictions, because they clarify the general connections between individual accuracy, diversity and group accuracy. These formal results show that diversity of perspectives/opinions will enhance group accuracy, and independence of group members will generally enhance diversity. At the same time, increasing individual accuracy will also enhance group accuracy. So, collective accuracy ("wisdom of the crowds") will benefit from communication wherever the benefits to individual accuracy exceed the reduction in diversity/independence. However, the formal results of this paper go beyond those insights in several key ways. First, we are not dealing with a collective voting or averaging case as in those "wisdom of the crowds" results. Instead, each piece of evidence is being weighted at its appropriate degree of reliability. Second, our results identify specific *structural constraints* on dependency relations (namely a unidirectional link between sources) that allow us to identify *a priori* when dependency will help and when it will hurt. In other words, we provide a first clarification of how the type of dependency relation itself matters. Of course, further work is encouraged to explore more complex forms of dependency relations (e.g., the incorporation of theory of mind relations between dependent sources).

By disentangling different information states (corroborative, but also contradicting and partial) and the structural representations of dependencies they may give rise to, we provide a new way of thinking about how dependencies fit within evidential reasoning. More precisely, we can first develop our understanding of the way in which lay reasoners infer the conditional influence of a dependency - as a function of source reliability, evidence-evidence and evidence-hypothesis structures, as well as various forms of background information. In so doing, we can

chart the otherwise nebulous territory of dependencies within evidential reasoning. But further, so as to complete this picture, the present formalism allows (via the application of probability theory) for the comparison of the products of lay reasoning (e.g., judgments and estimations) against an *informed* normative expectation – sidestepping the issue of the potentially intractable *general* normative account of dependence. In this way, we relax the ultimate normative constraint for approaching an understanding of dependence (i.e., the extrication of dependence from its context), so as to make meaningful in-roads into a psychological understanding of dependence - purposefully understanding dependencies (and their implications) *within* the contexts to which they are so inextricably bound.

Finally, we point to the implications of this work in applied domains of reasoning, including forensics, law, intelligence analysis, and medicine (although everyday reasoning is also intrinsically affected). First, where the use of formal approaches such as Bayesian Networks are feasible, this work makes a strong argument for the value of modelling dependencies in a careful and considered manner. Second, where such approaches are not readily applicable, our assumptions serve as a note of caution to drawing conclusions based on the naive intuition that dependencies are evidentially inferior.

ACKNOWLEDGEMENTS

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), under Contract [2017-16122000003]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. The research was also supported in part by the Leverhulme Trust under Grant RPG-2016-118 CAUSAL-DYNAMICS. The authors acknowledge and Agena Ltd for software support.

A subset of these results (Experiment 1) were presented at the Cognitive Science Society (Pilditch, Hahn, & Lagnado, 2018).

REFERENCES

Agena Ltd (2019). AgenaRisk (www.agenarisk.com)[Computer software].

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences*, 201615978.

Berg, S. (1993). Condorcet's Jury theorem revisited. *European Journal of Political Economy*, *9*(3), 437-446.

Berg, S. (1994). Evaluation of some weighted majority decision rules under dependent voting. *Mathematical Social Sciences*, 28(2), 71-83.

Bex, F. J., & Prakken, H. (2004). Reinterpreting arguments in dialogue: an application to evidential reasoning. *Legal knowledge and information systems. Jurix*, 119-129.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford University Press on Demand.

Clemen, R. T., Fischer, G. W., & Winkler, R. L. (2000). Assessing dependence: Some experimental results. *Management Science*, *46*(8), 1100-1115.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk analysis*, *19*(2), 187-203.

Condorcet, N. C. de. (1785). *Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix*, Paris. See I. McLean and F. Hewitt, trans., 1994.

Dawid, A. P., & Evett, I. W. (1997). Using a graphical method to assist the evaluation of complicated patterns of evidence. *Journal of Forensic Science*, *42*(2), 226-231.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*(1), 158.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive science*, *37*(1), 61-102.

Hahn, U., Harris, A. J., & Corner, A. (2016). Public reception of climate science: Coherence, reliability, and independence. *Topics in cognitive science*, *8*(1), 180-195.

Hahn, U. & Hornikx, J. (2016). A normative framework for argument quality: Argumentation schemes with a Bayesian foundation. *Synthese*, 193, 1833-1873.

Hahn, U., von Sydow, M. & Merdes, C. (2018) How Communication Can Make Voters Choose Less Well. *Topics in Cognitive Science*. 11, 194-206.

Harris, A. J., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The appeal to expert opinion: quantitative support for a Bayesian network approach. *Cognitive Science*, *40*(6), 1496-1533.

Højsgaard, S. (2012). Graphical independence networks with the gRain package for R. *Journal of Statistical Software*, *46*(10), 1-26.

Hogarth, R. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance, 21*, 40-46.

Hogarth, R. M. (1989). On combining diagnostic ‘‘forecasts’’: Thoughts and some evidence. *International Journal of Forecasting,* **5,** 593–597.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2.

JASP Team (2017). JASP (Version 0.8.4)[Computer software].

Jeffreys, H. (1961). *Theory of probability* (3rd Ed.). Oxford, UK: Oxford University Press.

Jönsson, M. L., Hahn, U., & Olsson, E. J. (2015). The kind of group you want to belong to: Effects of group structure on group accuracy. *Cognition*, *142*, 191-204.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal*, *7*(4), 317-337.

Ladha, K. K. (1995). Information pooling through majority-rule voting: Condorcet's jury theorem with correlated votes. *Journal of Economic Behavior and Organization*, 26, 353–372.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020-9025.

Madsen, J. K., Hahn, U., & Pilditch, T. D. (2018). Partial source dependence and reliability revision: the impact of shared backgrounds. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 722-727). Austin, TX: Cognitive Science Society.

Page, S. E. (2008). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies-New Edition*. Princeton University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2009). *Causality. Models, reasoning, and inference*. Second edition. New York: Cambridge University Press.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of experimental psychology*, *72*(3), 346.

Pilditch, T.D., Hahn, U., & Lagnado, D. (2018). Integrating dependent evidence: naïve reasoning in the face of complexity. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 884-889). Austin, TX: Cognitive Science Society.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive psychology*, *72*, 54-107.

Schum, D. A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review*, 105-151.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Northwestern University Press.

Smith, A. E., Ryan, P. B., & Evans, J. S. (1992). The effect of neglecting correlations when propagating uncertainty and estimating the population distribution of risk. *Risk Analysis*, *12*(4), 467-474.

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology*, *38*(2), 317-346.

Spellman, B.A. (2011). Individual reasoning. In B. Fischhoff & C. Chauvin (Eds.), *Intelligence Analysis: Behavioral and Social Scientific Foundations.* National Academies Press.

Wagenaar, W. A., Van Koppen, P. J., & Crombag, H. F. (1993). *Anchored narratives: The psychology of criminal evidence*. St Martin's Press.

APPENDIX

A. GENERAL NOTATION AND ASSUMPTIONS

To characterise cases of dependency advantages, we take the following models (as illustrated in Fig. A.1), wherein the sole difference between the left-hand (independent, $Model_I$) and right-hand (dependent, $Model_D$) is the influence of S_C on S_{B1}.

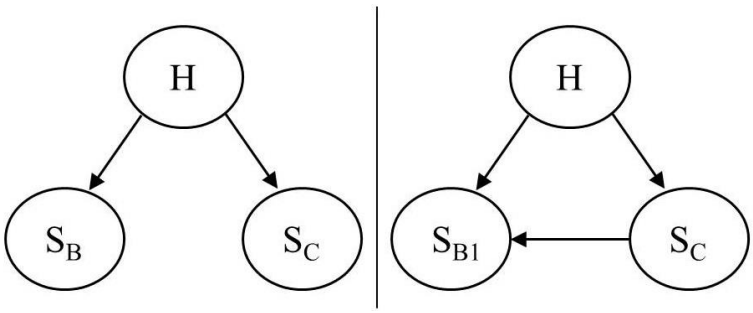

Figure. A.1. Graphical representation of a hypothesis (H) with two sources of evidence (S_B / S_{B1}, S_C) informing upon it. The left-hand model ($Model_I$) represents the independent case, whilst the right-hand model ($Model_D$) represents the dependent case.

For simplicity we are going to assume 'symmetry' of the error probabilities (so we will assume the true positive rates are the same as the true negative rates), as this characterizes both the experiments in the paper and the proof for the partial independence case below. However, the generalisation will be easy to see. So, in the two models we assume the following:

The conditional probability tables (CPTs) for S_C |H (in both models) and S_B |H in $Model_I$ is defined as:

Table. A.1. Conditional probability table (CPT) for the probability of false (bottom row) and true (top row) reports from sources, given the hypothesis is false (right-hand column) or true (left-hand column).

	H	$\bar{H}$
S	x	1-x
$\bar{S}$	1-x	x

The CPT for (S_{B1} | H, S_C) in $Model_D$ needs to consider not only H, but also the influence of S_C's (dis)agreement with H, and is therefore defined as:

Table A.2. Conditional probability table (CPT) for the probability of false (bottom row) and true (top row) reports from S_{B1}, given the hypothesis is false (right-hand pair of columns) and S_C (correctly) states it is false (right-most column) or S_C (falsely) states it is true (middle-right column), or, given the hypothesis is true (left-hand pair of columns) and S_C (falsely) states it is false (middle-left column) or S_C (correctly) states it is true (left-most column).

	H		$\bar{H}$	
	S_C	$\overline{S_C}$	S_C	$\overline{S_C}$
S_{B1}	u	v	1-v	1-u
$\overline{S_{B1}}$	1-u	1-v	v	u

A.1 Corroboration

To examine the conditions for dependency advantage in the case of corroboration, it is convenient to consider the odds form[13] of Bayes' rule:

Posterior Odds = Likelihood Ratio (LR) x Prior Odds

Considering the odds makes clear that in comparing dependent and independent corroboration we do not need to worry about the prior, as it is the same across both cases. So, quite simply, the dependent case will provide stronger corroboration when LR_D is greater than LR_I.

[13] The equation for posterior odds may be re-written as

$$\frac{P(H|E)}{P(\bar{H}|E)} = \frac{P(E|H)}{P(E|\bar{H})} * \frac{P(H)}{P(\bar{H})}$$

The odds form of Bayes' rule may also be converted back to posterior probabilities via

$$P(H|E) = \frac{Posterior\ Odds}{(Posterior\ Odds + 1)}$$

Given that we have two pieces of evidence (e.g., one report from Bailey and one from Campbell), the relevant LRs are those of the *two pieces of evidence presented together*. This is the **product** of the two individual LRs.

In the independent case this is just (assuming the notation above):

$$\frac{x}{(1-x)} * \frac{x}{(1-x)}$$

For the dependent case, we have (again with notation above)

$$\frac{u}{(1-v)} * \frac{x}{(1-x)}$$

So, all we need to know to assess the corroborating case is whether

$$\frac{u}{(1-v)} > \frac{x}{(1-x)}$$

i.e., whether the diagnosticity of Bailey has increased as a result of the dependency.

In the main text above and in our consideration of the partial information case below, we focus on the consequences of the plausible assumption that a correct report from Campbell makes Bailey more accurate, whereas an incorrect one reduces her reliability, i.e.

$$u > x > v$$

in this case, whether it does or does not involves two opposing forces:

- How much bigger is *u* relative to *x* (i.e., how much does it help make B more accurate when C is correct) (bigger numerator will increase LR for Bailey)

and

- How much smaller is v than x (i.e., how much does it HURT B when C is wrong) (because the smaller v, the bigger (1-v)…. (bigger denominator will decrease LR for Bailey)

The winner between these two opposing forces determines whether corroboration helps or hinders.

For a visualisation of the above constraints in action, see also Appendix B.

A.2 Contradiction

In the contradicting case, we have the same thing, except now we have, for the independent case:

$$\frac{x}{(1-x)} * \frac{(1-x)}{x}$$

And for the dependent case:

$$\frac{(1-u)}{v} * \frac{(1-x)}{x}$$

And whether there is an advantage or a disadvantage is determined by whether u and v are such that they make B **less diagnostic** than C.

A.3 Partial Information

In the partial information case, we have the same notation and assumptions as in A.1, including 'symmetry' of the error probabilities, and that both true positives and true negatives are higher than 50% (i.e., assumption 1; $x > 0.5$).

Following from Table A.2, for this case we therefore assume:

$$u > x > v > 0.5$$

This inequality is a summary of *assumption 1* (all sources are generally accurate) and *assumption 2* (correctly provided information is helpful), where,

$$u - x > x - v$$

Such that *assumption 3*, the 'boost' in accuracy when provided second-hand information is correct, is greater than the 'drop' in accuracy when provided second-hand information is

incorrect. This assumption is necessary for this proof as otherwise the hypothesis is not always true. **NB**: In the experimental studies, assumption 3 is not necessary to induce a dependency advantage, this is because the starting accuracy of sources is generally high (0.8). As starting accuracies approach the 0.5 threshold, assumption 3 becomes more and more necessary.

B. VISUALISATION

Next, we present a visualisation of how dependency gives rise to advantage and disadvantage across regions of the parameter space in Fig. B.1 below. The color map represents the degree of difference between dependent and independent posterior: differences > 0 reflect a dependency advantage; differences < 0 a dependency disadvantage. These differences are shown across all possible values of u and v (the conditional reliability of Bailey when Campbell is correct, and when Campbell is incorrect, respectively), for three different levels of independent accuracy x. Row 3 corresponds to the independent source reliability of .8 used in Exp.1 and 2. Hence mapping the distribution plots of Fig. 6 and Fig. 9 main text to the u and v axes will give a sense of the outcomes in the BIBNs for those studies.

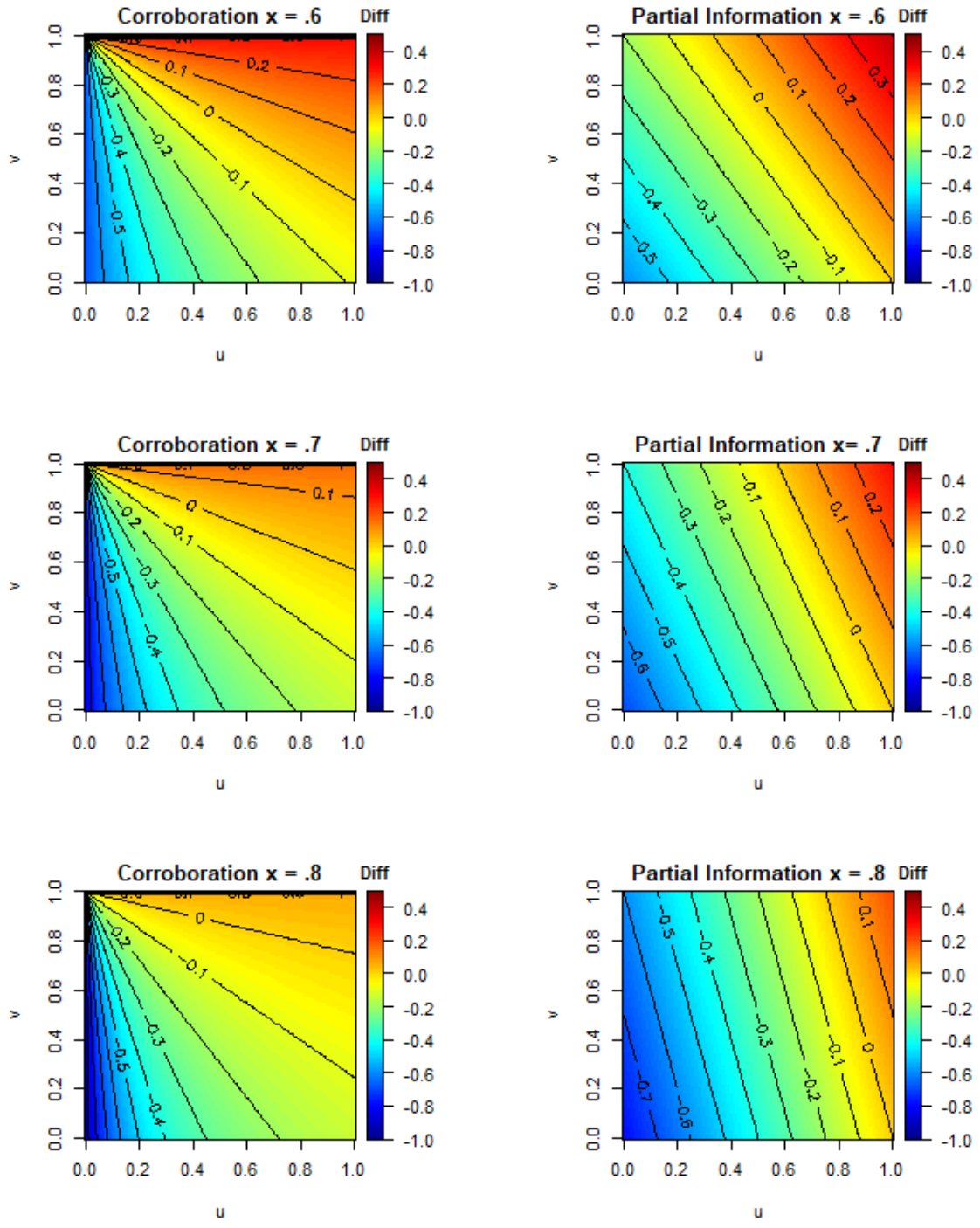

Fig. B.1 The plot visualises regions of dependency advantage (Diff > 0) and dependency disadvantage (Diff < 0) relative to independent evidence for the corroboration (left panels) and partial information case (right panels). X represents the accuracy of the independent source (see A.1 above), u represents the reliability of the dependent source where the second report is correct, v represents the reliability where that report is false (see Table A2 above). Prior = 5 in all cases. Row 3 corresponds to the conditions of Exp. 1 and 2.

C. PROOF

Finally, with the above assumptions, to demonstrate and advantage of the Model_D over Model_I, we have to prove the following:

$$P(H|S_B) < P(H|S_{B1})$$

By Bayes theorem this is equivalent to proving:

$$\frac{P(S_B|H) \times P(H)}{P(S_B)} < \frac{P(S_{B1}|H) \times P(H)}{P(S_{B1})}$$

Which is equivalent to proving

$$\frac{P(S_B|H)}{P(S_B)} < \frac{P(S_{B1}|H)}{P(S_{B1})} \qquad (A.1)$$

We note here that in proving (A.1), we ensure the requirement that $P(S_B|H) < P(S_{B1}|H)$ *always holds*, given our assumptions. Again, to avoid a massively complex proof, we make the simplifying assumption that $P(H) = 0.5$. Given this, the LHS of (A.1) is straightforward to compute:

$$\frac{P(S_B|H)}{P(S_B|H)P(H) + P(S_B|\bar{H})P(\bar{H})} = \frac{x}{x\frac{1}{2} + (1-x)\frac{1}{2}} = 2x \qquad (LHS)$$

The RHS of (A.1), however, is more complex because we have to incorporate the conditional dependency:

$$\frac{P(S_{B1}|H,S_C)P(S_C|H) + P(S_{B1}|H,\overline{S_C})P(\overline{S_C}|H)}{P(S_{B1}|H,S_C)P(S_C|H)P(H) + P(S_{B1}|H,\overline{S_C})P(\overline{S_C}|H)P(H) + P(S_{B1}|\bar{H},S_C)P(S_C|\bar{H})P(\bar{H}) + P(S_{B1}|\bar{H},\overline{S_C})P(\overline{S_C}|\bar{H})P($$

$$= \frac{ux + v(1-x)}{\frac{1}{2}(ux + v(1-x) + (1-v)(1-x) + (1-u)x)}$$

$$= \frac{ux + v(1-x)}{\frac{1}{2}} = 2\big(ux + v(1-x)\big) \qquad (RHS)$$

Now, considering LHS and RHS, it follows from (A.1) that we have to prove:

$$x < ux + v(1-x)$$

Or, equivalently, that

$$x - ux + vx - v < 0 \qquad (A.2)$$

To prove (A.2), we note that

$$x - ux + vx - v$$

$$= x(1 - u + v) - v$$

$$< \frac{1}{2}(u+v)(1-u+v) - v$$

(the above is a consequent of *assumption 3,* $u - x > x - v$, which implies that $x < \frac{1}{2}(u+v)$)

$$= \frac{1}{2}(u - u^2 + uv + v - uv + v^2) - v$$

$$= \frac{1}{2}\big(u(1-u) + v(v-1)\big) \qquad (A.3)$$

So to prove (A.2) we now only need to prove that (A.3) >= 0, which is the same as proving that:

$$u - u^2 < v - v^2 \qquad (A.4)$$

But, we know that $u > v > 0.5$. Hence, we can assume that

$$v = 0.5 + \alpha \text{ (where } 0 < \alpha \leq 1)$$

and that

$$u = 0.5 + \alpha + \beta \text{ (where } 0 < \alpha + \beta \leq 1)$$

Substituting into the LHS of (A.4) we get:

$$(0.5 + \alpha + \beta\ ) - (0.5 + \alpha + \beta\ )^2$$

$$= 0.5 + \alpha + \beta$$

$$- (0.25 + 0.5\alpha + 0.5\beta + +0.5\alpha + \alpha^2 + \alpha\beta + 0.5\beta + \alpha\beta + \beta^2)$$

$$= 0.25 - \alpha^2 - 2\alpha\beta - \beta^2$$

We then get the RHS of (A.4):

$$0.5 + \alpha - (0.5 + \alpha)^2 = \ 0.5 + \alpha - (0.25 + \alpha^2 + \alpha)$$

$$= 0.25 - \alpha^2$$

Finally, by comparing the LHS and RHS of (A.4) and noting that α and β are both positive, it follows that the inequality (A.4) is proven. *Quod Erat Demonstrandum*.